# Computational analysis of the *Caenorhabditis elegans* genome sequence.

## Thesis

UNRESTRICTED

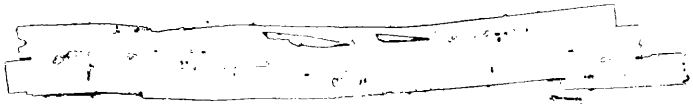
# Computational Analysis of the *Caenorhabditis elegans* Genome Sequence

Steven John Mathias Jones, BSc, MSc

A thesis submitted in partial fulfilment of the requirements of the Open University for the degree of Doctor of Philosophy

June 1999

The Sanger Centre



DATE OF AWARD: 20 DECEMBER 1999

# CONTENTS

# LIST OF TABLES

Table 5-2 Proportion of tandem and inverted gene pairs in the genetic compartments.

# LIST OF FIGURES

Figure 5-4 Similarity of inverted gene pairs.

# Abstract

The genomic sequencing of the model genetic organism, the nematode *Caenorhabditis elegans* is now essentially complete, representing the first genome sequence to be derived for a multicellular organism. This thesis describes the strategies and software tools that have been utilized in the analysis of the genomic sequence. Preliminary analysis of genomic organisation is also presented.

*C. elegans* chromosomes do not store genetic information in a uniform manner. Gene density varies between different chromosomal regions and between chromosomes. The highly recombinagenic autosomal arms possess more repetitive elements and generally have a lower gene density than the recombinationally suppressed central regions. Although, the gene density within autosomal arms is higher than had been previously expected. A positive correlation is observed between the number of genetically defined loci from a chromosomal region and the expression rate of a region as estimated by the abundance of Expressed Sequence Tags (ESTs). A similar positive correlation is observed with the proportion of genes possessing similarity to non-nematoda proteins. Chromosomal regions with a high density of gene clusters have fewer genetically derived loci. Demonstrating that redundancy reduces the genetic accessibility of a region towards classical genetic approaches.

Introns are larger on the autosomal arms than the central clusters. Exon length shows no correlation with chromosomal position but increases with expression rate. Stop codon preference is also influenced by expression rate.

Clusters of similar genes are also found on the *C. elegans* chromosomes although their distribution is not random. The majority of gene clusters have been determined to lie on chromosome V and the left arm of II. The orientation of the genes within gene clusters suggests that inversion events are common and provide a selective advantage. Alternative splicing has also been studied and the results suggest that many alternative transcripts can be attributed to errors in splice acceptor processing.

# General Introduction

As a model organism *C. elegans* has proven to be particularly amenable to genetic analysis. The ease with which it can be cultured in the laboratory and its relatively short generation time may have been some of the more practical reasons for it initially being chosen by Sydney Brenner as a model organism [Brenner 1974]. But many more compelling reasons for its study exist. Its invariant lineage coupled with the fact that it is transparent at all life stages has made *C. elegans* an extremely powerful tool in the study of developmental biology and increasingly in the study of apoptosis [Metzstein *et al.* 1996] and longevity [Barnes *et al.* 1997]. The rudimentary nervous system comprising of only 302 neurons in the adult hermaphrodite has also made *C. elegans* influential in the fields of behavioral research and neurophysiology.

The construction of a clone based physical map over several years [Coulson *et al.* 1986, 1988, 1991 and 1995] provided the enabling resources to attempt in 1990 initial pilot studies investigating the feasibility of complete genomic sequencing of what was estimated to be a 100 megabase genome. The success of the early pilots [Sulston *et al.* 1992; Wilson *et al.* 1994] led to the completion of the entire genomic sequence by the original international partnership consisting of the Sanger Centre, Cambridge, UK and the Genome Sequencing Center, St. Louis MO, USA. [*C. elegans* Sequencing Consortium 1998]. The estimated error rate of the sequence being estimated to be better than 1 error in 10,000 base pairs. The

complete genomic sequencing of *C. elegans* represents the first genome of a multicellular organism to be derived.

Prior to the completion of the *C. elegans* genome the complete genomic sequence of a number of single celled organisms had been derived. The first bacterial genome to be sequenced was that of *Haemophilus influenzae* [Fleischmann *et al.* 1995] which adopted a whole shotgun strategy and automated sequencing approaches. The sequencing of the *Helicobacter pylori* genome eclipsed the more established genome project of *Escherichia coli* [Blattner *et al.* 1997], which failed to embrace the most up to date sequencing technology available. In total, over twenty bacterial genomes have now had their complete genomic sequence derived. Most notably those of *Helicobacter pylori* [Tomb *et al.* 1997; Alm *et al.* 1999], *Mycobacterium tuberculosis* [Cole *et al.* 1998] and *Bacilllus subtilis* [Kunst *et al.* 1997]. Many more bacterial genomes are in progress, for example those of *Yersinia pestis*, *Streptomyces coelicolor*, *Salmonella typhi* and *Pyrococcus furiosus*.

The sequencing of the first eukaryotic chromosome was completed in 1992 with the publication of chromosome III from the yeast *Saccharomyces cerevisiae* [Oliver *et al* 1992]. This was followed by the entire genome of *S. cerevisiae* in 1996, a task ultimately involving an international consortium consisting of over 600 scientists from Europe, North America and Japan [Goffeau *et al.* 1996].

Gene detection and annotation has been relatively easy in the single celled organisms primarily because the general lack of intronic sequence means that genes can be detected as large open reading frames and because many of the genes had already been independently studied. The latter reason being particularly relevant for *E. coli* and *S. cerevisiae*. However, merely relying on the

presence of statistically significant open reading frames is problematic. A reappraisal of the gene predictions on chromosome III of *S. cerevisiae* by Koonin *et al.* 1994 determined a case where a large ORF was erroneously assigned as a gene instead of two smaller ORFs present on the opposite strand. This study also determined that smaller exons had also been missed from some rare genes in *S. cerevisiae* which possess introns. More complex approaches have also been used to discriminate between non-coding and coding sequence using hidden markov models in bacterial sequence such as that of *E. coli* and *H. influenzae* [Borodovsky *et al.* 1994a,b].

In more complex multicellular eukaryotes analysis remains more challenging. The presence of introns and a general trend of lowered gene density in more complex organisms hamper gene prediction. Comparative analysis with large amounts of annotated genomic sequence from other metazoans has not been possible for *C. elegans* for almost the entire duration of the sequencing project. Until recently most of the human genome lacked a coherent sequencing program and sequencing efforts concentrated mainly around disease loci. In addition, genomic sequencing of the mouse genome remained in what can be considered pilot comparative studies [Koop *et al.* 1994a,b]. Other efforts to provide sequence data for comparative analysis during this period were also proposed, most notably that of the pufferfish *Fugu rubripes* [Brenner *et al.* 1993].

*Fugu* has a smaller genome primarily due to fewer repetitive intronic and intergenic sequences whilst it maintains a similar complement of genes to other vertebrates. Sequence derived from its 400 megabase genome was proposed primarily as an aid to the study human and other vertebrate sequences. However, readily available sequence would be useful for the study of all metazoans. Since

the initial publication of the utility of *Fugu,* interest in its complete genomic sequencing has waned and recent developments indicate that funding is being directed towards the genetically amenable mouse to provide a second vertebrate genome after that of the human.

Recently a new impetus in genomic sequencing has provided the means to derive draft sequences for both *Drosophila melanogaster* [Rubin 1998] and human by early 2000 [Waterston and Sulston 1998, Wadman 1999]. This will provide *C. elegans* with a number of homologues which can be used to refine existing gene predictions. However, to be able to refine the predictions of almost all *C. elegans* genes, the sequence of a closely related nematode will be required. This is being addressed through the sequencing of the nematode *Caenorhabditis briggsae.* This nematode is morphologically similar to *C. elegans*, has a similar life cycle and is thought to have shared a common ancestor with *C. elegans* less than 20 to 30 million years ago. The divergence between these two organisms is such that functional elements remain conserved and readily detectable. Non-functional DNA sequences bear little similarity, whilst the extensive synteny between the two species allows the rapid and unambiguous assignment of homologous genes. Almost 10% of the genome of *C. briggsae* has so far been sequenced (M. Marra, unpublished results).

This thesis discusses the tools and methodologies employed in the analysis of the genomic sequence data and attempts to examine some of the aspects by which genetic information is organized within the *C. elegans* chromosomes. The thesis is divided into five sections each with its own introduction, methodology results and discussion.

The first section describes how the analysis of the *C. elegans* genome has taken place during the sequencing project. The software tools used to accomplish this task are also outlined. Analysis and annotation strategies are also described as well as the future directions that analysis of this genome may ultimately take. The second section investigates how information is organized within the *C. elegans* six chromosomes. This encompasses some of the differences which are observed in gene structure and expression between different chromosomal regions. The amenability of various chromosomal regions to classical genetic analysis is also explored. In addition, the distribution of putative homologues to both human and yeast genes is also explored as well as the distribution of repetitive elements. The third section investigates the effect of the transcriptional rate on genomic features using the abundance of Expressed Sequence Tags (ESTs) as a measure transcriptional activity. The effects of transcription rate on intron and exon size, stop codon preference and synonymous codon bias is explored. The forth section investigates the types of alternative splicing observed within *C. elegans*. The fifth section investigates the presence of closely related groups of genes known as gene clusters and their chromosomal distribution. In addition, some of the mechanisms by which the gene clusters may be formed and maintained are also investigated.

# Methodology in *C. elegans* Sequence Analysis

*Introduction*

In the construction of the sequence map of *C. elegans* various computational tools and strategies were developed to allow the high throughput, collection, assembly and management of sequence data [Dear *et al.* 1998; Wendl *et al.* 1998]. In tandem with this effort computational strategies were also implemented to allow the high throughput biological analysis of the finished genomic sequence data. This section discusses the strategies, approaches and computational tools that were implemented in the analysis of the *C. elegans* sequence.

The sequencing of the *C. elegans* genome has relied almost entirely on the sequence ready clones provided by the physical map [Coulson *et al.* 1986, 1988, 1991 and 1995]. Initially, only cosmid clones were selected and sequenced from the central regions of the autosomes and the entire X chromosome. However gaps in cosmid coverage still remained in the physical maps derived from these regions. In addition, physical maps derived from the autosomal arms had much poorer cosmid coverage. Therefore, in the latter stages of the project techniques and protocols were derived to enable the sequencing of PCR products, fosmids and YAC clones which were used to enable sequence contiguation. The resulting product has been a set of overlapping DNA sequences derived from physical map clones. This approach has provided the advantage that as almost every sequence is derived from a discrete clone, an obvious link is established between the

sequence data and the physical map. In addition, as the average size of sequence resulting from each cosmid is around 30KB and few YAC derived segments exceed 100Kb, sequences from the *C. elegans* sequencing project remain a manageable size for the majority of sequence analysis tools. Also, by submitting finished sequences to the databanks on a clone by clone basis, finished regions are submitted to the public databanks rapidly rather than waiting for larger sequence contigs, such as chromosomes, to be completed.

It is important to note that for each clone the resulting sequence will not necessarily represent its entire insert. This is because the clones chosen will usually have a significant overlap and submitting the entire overlap region would result in a large amount of redundant sequence within the public sequence databases. Therefore sequences are clipped to provide between 100 and 200bp of overlapping sequence between contiguous clones. The overlapping sequence in part ensures that significant matches from homology searches are not lost at the ends of the sequences and also that the overlapping sequence provides an obvious link between sequences determined from overlapping clones.

The positions of the cloning sites are recorded, where possible, so that in the future complete insert sequences can be derived for restriction digest analysis and for interpreting transformation rescue experiments.

*The Database*

It is unsurprising that the major problem in sequence analysis of the *C. elegans* genome has been that our understanding of eukaryotic genomes is far from complete. Therefore, in order to store the *C. elegans* DNA sequence along

with the resultant analyses and annotation the database used was required to be as flexible and configurable as possible. This would allow new methods and data to be easily incorporated into the database schema and graphical displays. Another requirement was that the database should also allow the rapid editing of predicted gene structures and genome features when more information becomes available i.e. approaches to allow the easy visualization of genomic features needed to be established. In addition, the database also needed to be easily disseminated so data can be readily accessible by the end user biologist. The database ACEDB has been developed primarily by Richard Durbin and Jean Thierry-Mieg [1991-] to provide such functionality for the *C. elegans* genome project and its flexible approach has meant that it has at various times been adopted by several other genome projects.

*The Dataflow*

When a sequencing project for a clone is finished the required sequence is excised from the GAP database [Bonfield *et al.* 1995] using the program MKCON-GAP [Unpublished, available from http://www.sanger.ac.uk/Software]. At this point a set of automated analyses are performed on the DNA sequence and the results converted into the ACEDB file format. A general overview of the analysis process and the tools employed is provided in figure 1-1. Each analysis program is called from within a UNIX shell script. The script also logs each process as well as its exit status so that any problems can be quickly identified and resolved. When automated analysis is complete the resulting log file is mailed to the initiating user.

The GENEFINDER algorithm was invoked manually from within ACEDB to predict genes present across sequence boundaries.

A brief summary of the tools utilized is provided in table 1-1.

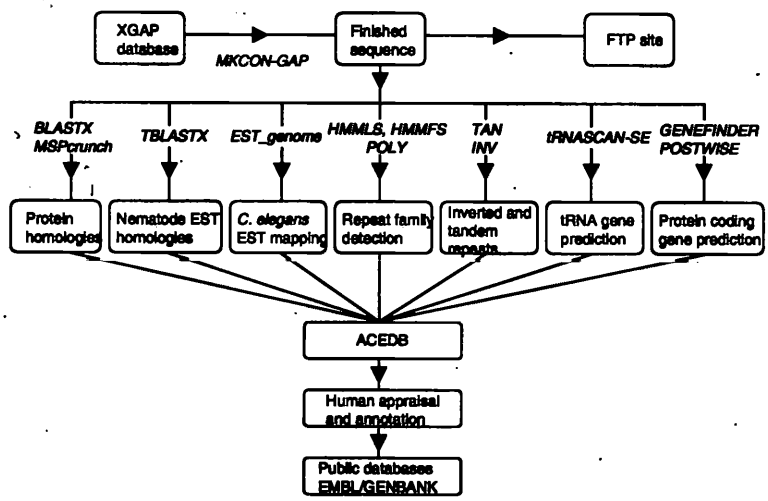



Figure 1-1: Sequence Analysis Overview.

*Genefinding*

The program GENEFINDER (Green and Hillier unpublished) was developed to predict putative protein coding genes within the *C. elegans* sequence data. GENEFINDER uses statistical criteria derived from log likelihood ratios to detect potential genes based on genomic features such as splice sites, translation start

sites and codon biases. A dynamic programming algorithm is used to find a set of non-overlapping candidate genes with the highest total score for each DNA strand.

Table 1.1. Summary of analysis programs employed

| | |
|---|---|
| GENEFINDER | *ab-initio* gene prediction [Green and Hillier unpublished]. |
| POSTWISE | Gene Prediction based on protein homology [Birney 1997]. |
| tRNASCAN-SE | transferRNA gene prediction [Lowe and Eddy 1997] |
| INV | Inverted Repeat detection [R. Durbin unpublished] |
| TAN | Tandem Repeat Detection [R. Durbin unpublished] |
| HMMLS, HMMFS | Hidden Markov Model detection of repeat families [Eddy 1995-] |
| POLY | Detection of repeat family members present in tandem arrays [R. Durbin unpublished] |
| MSPcrunch | BLAST Post Processor [Sonnhammer and Durbin 1994] |
| BLASTX | Six frame translation and comparison to protein database [Altschul *et al.* 1990] |
| TBLASTX | DNA vs. DNA comparisons at protein level [Altschul *et al.* 1990] |
| EST_genome | Alignment of EST sequences to Genomic DNA [Mott 1997] |

One of the major problems in *ab-initio* gene prediction methods is the inability to accurately and sensitively detect splice site signals. Splicing is extensive in *C. elegans* with >96% of transcripts predicted to possess more than one exon (figure 1-2). The various methodologies developed to determine intron splice sites, including those used in GENEFINDER, produce high enough numbers of false positive splice sites to make gene finding problematic.

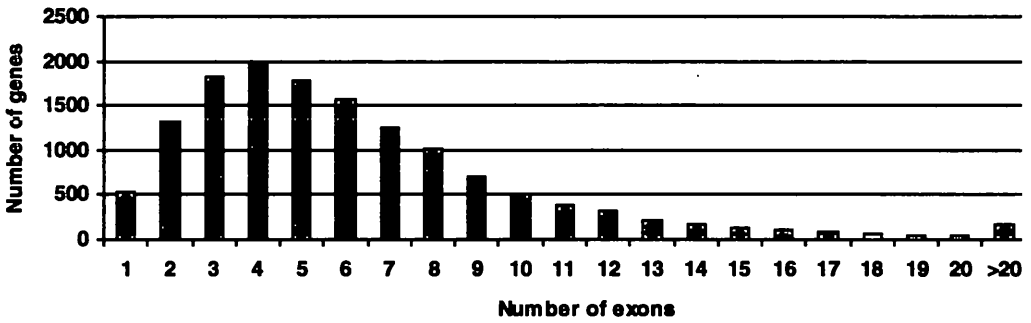



Figure 1-2. Variation of predicted exon number per gene. Data collected from 14,251 predicted genes.

*C. elegans* also lacks a detectable branch site consensus, in contrast to both yeast and mammalian introns. However gene detection is aided by the fact that *C. elegans* introns tend to be relatively small with an overall median size 57 base pairs and also because around 25% of the total genomic sequence is predicted to be protein coding. The translation initiating AUG codon does have a weak surrounding consensus, showing a preference for A in the preceding four bases; however the 5' ends of genes remain difficult to predict with high accuracy. One potentially confusing aspect of gene prediction is *trans*-splicing in *C. elegans* [Krause and Hirsh, 1987]. In this process short RNA leader sequences are spliced onto the 5' ends of mRNAs. The recognition site for this splice leader sequence addition has the same consensus sequence as the standard *cis*-splice acceptor site. This can lead to the gene prediction misinterpreting the initiating exon as an internal exon and the gene being erroneously extended upstream. Another problem is the fact that some genes in *C. elegans* are co-transcribed producing polycistronic messages [Spieth *et al.* 1993]. One feature of such operons is that

the distance between the polyadenylation site of an upstream gene and the *trans*-splice site of the downstream gene is short, usually being in the region of 100 base pairs. Together the close nature of the genes and *trans*-splicing signals create the strong possibility that many operons will be mis-predicted as a single gene.

As the genome project has proceeded the authors of GENEFINDER have been able to utilise the increasing amount of genome data for training and calibration as well as improving their computational methods. The accuracy of the current version of GENEFINDER in exon prediction as well as the previous implementation is shown in table 1-2. Using an experimentally confirmed set of gene structures 85% of GENEFINDER exon predictions were found to be correct. GENEFINDER is therefore an effective method in determining at least part of most genes. As most genes in *C. elegans* have 4 or more exons [figure 1-2] the proportion of gene predictions which are 100% correct using GENEFINDER alone will be significantly less than 85%. Particularily problematic to predict are the initial and terminating exons as they possess a single intron splice site.

As gene finding methodologies are currently non-optimal, consolidation of gene predictions with other biological information has proven to be essential. The most informative sources of extra information have been expressed sequence tags (ESTs) from *C. elegans* [McCombie 1992; Waterston *et al.* 1992; Kohara 1996] and similarity to protein sequences present in the public databases. Therefore many of the initial GENEFINDER predictions have been manually edited to consolidate the EST and protein homology information as well as information provided by independent *C. elegans* investigators. The ACEDB sequence display

(FMAP) allows both the visualisation of the homologous regions and the rapid editing of the gene structures.

Table 1-2: GENEFINDER appraisal.

| Version | Exons | Exact match | | | Overlap | | | 5' | | | 3' | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | N | Acc | Cov | N | Acc | Cov | N | Acc | Cov | N | Acc | Cov |
| To 1/5/98 | 285 | 219 | 0.77 | 0.84 | 250 | 0.88 | 0.95 | 230 | 0.81 | 0.88 | 238 | 0.84 | 0.91 |
| 1/5/98 on | 285 | 241 | 0.85 | 0.92 | 259 | 0.91 | 0.99 | 249 | 0.87 | 0.95 | 250 | 0.88 | 0.95 |

The test set consisted of 27 genes where their structure had been confirmed experimentally. The 27 confirmed genes contained in total 262 confirmed exons. Half the intergenic distance between the neighbouring genes was taken and the sequences concatenated in random orientation to simulate a single large sequence contig. The size of the sequence contig was 199kb. The identity of the genes used was unknown to the GENEFINDER authors. Acc refers to the proportion of predicted exons correct. Cov refers to the proportion of true exons predicted correctly. Overlap refers to predictions which at least overlap with a correct exon. 5' and 3' refer to exon predictions where the respective ends are predicted correctly. The GENEFINDER appraisal was carried out using the program gff_predict_ana.pl [T. Hubbard, unpublished].

It is envisaged that future versions of GENEFINDER will be able to make use of homology data to protein and EST sequences directly as part of the initial gene prediction. This approach should eliminate much of the requirement for the manual editing of gene structures

ESTs provide valuable transcriptional data, not only confirming that predicted genes are indeed transcribed *in-vivo* but also indicating the position of intron/exon boundaries allowing a transcriptional map of the genomic sequence to be determined. Although the EST datasets have been subjected to normalisation techniques, (in the case of *C. elegans* ESTs this has involved the use of use of abundant cDNA species as hybridisation probes so that these genes can be eliminated from further selection for tag sequencing) EST datasets can still also provide some information pertaining to the relative expression rate of genes.

EST mapping to genomic sequence is carried out by the program EST_genome [Mott 1997]. This program aligns the EST sequence to the genomic sequence while preferentially allowing gaps at intron (NN/GT..AG/NN) boundaries. The resulting alignments therefore reflect closely the intron/exon structure of the genes from which they are transcribed. A wrapper for the EST_genome program was developed called EST_genome2ace. This PERL program converts the output of EST_genome into ACEDB format and also records in ACEDB format the co-ordinates of the introns which are confirmed by the EST data. This not only allows for the construction of confirmed intron splice site datasets for those interested in refining gene prediction algorithms but more importantly enables the automated identification of incorrect gene predictions. This latter aspect has proven to be particularly important when EST projects and genomic sequencing are proceeding concurrently, as ESTs for a gene will often be derived after the initial gene prediction has been created.

A small fraction of *C. elegans* introns (<1%) begin with GC instead of the canonical GT and at present can only be detected reliably where confirmatory EST data exits. Although work has shown that aberrant 3' splice sites which lack the canonical AG sequence can still splice at the normal site, [Aroian *et al.* 1993; Zhang and Blumenthal 1996] only one putative non-AG 3' splice site has been found in wild-type sequence (EMBL accession number Z92828).

*Protein similarities*

Protein homology mapping is done by comparing a six-frame translation of the genomic sequence to a protein database using BLASTX [Altschul *et al.* 1990]

in conjunction with a BLAST post-processor, MSPcrunch [Sonnhammer and Durbin 1994]. MSPcrunch improves the signal to noise ratio by the elimination of matches due to low complexity regions whilst increasing the significance of low scoring fragmentary hits to the same protein sequence. MSPcrunch also has the added advantage of being able to produce output in the ACEDB file format.

It is important to note that when searching using large genomic DNA sequences that the maximum number of alignments reported in the BLAST output is increased by making the B parameter large (e.g. >=1,000,000), otherwise some *bona-fide* matches may not be reported. MSPcrunch also limits, by default, the number of aligned matches reported for any region to 20. This prevents the storage of excessive amounts of data, for example in regions where genes are strongly conserved and extensively studied in many different species e.g. histone genes.

A non-redundant protein database, SWIR [E. Sonnhammer and P. Rice, unpublished], has been maintained for protein homology searches. The SWIR database is made up from the three protein databases WormPep [*C. elegans* Sequencing Consortium 1991-], SwissProt and Trembl [Bairoch and Apweiler 1997]. Duplicate sequences in these databases are detected and removed by their database cross-references and peptide sequences are retained in the priority of Wormpep, SwissProt and then Trembl. Protein sequences which are truncation or fragments of larger protein sequences are also removed. Redundancy of this protein database is limited further by the removal of closely related sequences [P. Rice unpublished]. This is achieved by first identifying candidates by their dimer composition and eliminating sequences where the identity is 95% or greater as determined by a Needleman-Wunsch alignment [Needleman and Wunsch 1990].

The ability to searching a non-redundant database as opposed to searching the complete constituent databases is that the computational time required is greatly reduced whilst not compromising the subsequent analysis.

Even with the use of non-redundant protein datasets maintaining up to date protein homologies for genomic sequence remains an increasingly computationally demanding task. Therefore, it is necessary to implement incremental updates of protein homologies. This is aided by the fact that each SWIR release is accompanied by datasets of proteins which have changed, have been deleted and are new since the last release. By removing references to deleted and changed proteins from the ACEDB database and by searching the entire genomic data against only those protein sequences which are new or have changed the protein homology data in ACEDB can be relatively rapidly updated.

Another source of protein coding elements can be found from EST datasets. To exploit this data, the genomic sequence is used to search a dataset consisting of ESTs derived from other nematode species. Comparisons are done at the protein level using TBLASTX, which compares conceptual six-frame translations of both the EST sequences and the *C. elegans* genomic sequence. Currently, the EST dataset consists predominately of ESTs derived from *Brugia malayi* (the causative agent of elephantiasis) and *Caenorhabditis briggsae*. ESTs are also represented from a variety of nematodes such as *Onchocerca volvulus* (causitive agent of river blindness), *Toxocara canis* (canine roundworm), *Haemonchus contortus* (sheep nematode parasite) and *Strongyloides stercoralis* (human nematode parasite).

The searching of ESTs derived from within the nematoda should allow detection of gene families that have arisen solely within the nematode lineage or

undergone rapid evolutionary change within the nematode lineage. The ability to use EST data in this case is pertinent as very few protein sequences from nematoda other than *C. elegans* presently exist in the public databases. As the EST dataset was relatively small and because ESTs from other nematodes were continually being deposited into dbEST [Boguski *et al.* 1993] the rate of gene discovery within these other nematodes was high. Since the primary utility of these data was to aid gene prediction and annotation in *C. elegans*, the EST database was remade automatically each week incorporating all available EST sequences from dbEST using the Sequence Retrieval System [Etzhold *et al.* 1996].

As the public protein database become more complete, it also becomes more feasible to predict genes based on homology information. Gene predictions based on homology data are produced using POSTWISE [Birney *et al.* unpublished]. POSTWISE uses an algorithm that combines gene prediction and protein homology in a single probabilistic model [Birney 1997]. The POSTWISE algorithm is used to predict genes as part of the initial DNA sequence employed in the *C. elegans* project. Using POSTWISE allows the first pass annotation process to be expedited as intron/exon structures based on protein homology are presented to the annotator for appraisal.

It is envisaged that tools such as POSTWISE will play an important role in the curation and updating of the predicted gene structures. As new orthologs and paralogs enter the protein databases, POSTWISE predictions can be automatically compared with the actual predictions and conflicts can easily be flagged.

Currently the *ab-initio* detection of non-protein coding genes has been limited to prediction of tRNA genes. This has been done using tRNAscan-SE [Lowe and Eddy 1997]. This approach utilises the program tRNAscan [Fichant and Burks 1991] to rapidly identify initial tRNA gene candidates. The resultant set is filtered using the COVE probabilistic RNA prediction package [Eddy and Durbin 1994]. The advantage of the COVE post-processor is that many of the tRNA gene false positives from tRNAscan can be identified and removed. The post-processor is also able to predict the amino acid charged by the tRNA gene. In addition, genes that are predicted to possess incomplete primary or secondary structure are marked as pseudogenes.

*Repetitive Sequences*

As with mammalian genomes, repetitive elements in *C. elegans* also represent a significant fraction of the genome. The repetitive elements are classified into two types, local and dispersed repeat family elements.

Local repeat families usually consist of unique sequence either repeated tandemly or as an inverted repeat and are detected using the programs INV and TAN [R. Durbin unpublished]. Of the dispersed repeats, 38 families have been determined [Durbin unpublished results; Lewis and Eddy unpublished], although this analysis is far from complete and more families will be derived [Holmes and Durbin unpublished]. Many of the dispersed repeat elements are related to transposable elements [Smit 1996]. It is likely that many repeat elements are representative of the remnants of transposable elements which are now extinct within the *C. elegans* genome. Most of the repeat families are described by a hidden markov model and are detected using the HMMER software [Eddy 1995-].

Not all repeat sequences are detected using probabilistic models; for example, CeRep25 which is a palindromic sequence and CeRep26 which is a tandem array of the telomeric repeat TTAGGC are both identified by the program POLY [R. Durbin, unpublished]. PERL scripts are used to convert the output of the repeat searching programs employed into ACEDB format.

*Protein Annotation*

Although heavily studied as a genetic model organism only a small fraction (approximately 4%) of the genes found in the sequence can be identified as having been previously studied experimentally by independent researchers. As the sequencing of the *C. elegans* genome itself was a major undertaking, an additional systematic biochemical study of the predicted genes could not practically be carried out in tandem with the sequencing project. Therefore almost all the functional information acquired for the predicted protein products has been inferred computationally, primarily derived from protein similarity data. Although such a strategy is beset with many caveats, the approach has been well received by the *C. elegans* community, allowing at least a putative function to be attached to many predicted genes in the *C. elegans* database.

One problem in the assignment of putative function is the variance between different human annotators. Annotation may vary in detail, nomenclature, accuracy and spelling. A common error is 'over annotation' where the protein homology is over interpreted, resulting in the annotation being more specific than the homology alone justifies. For example, a protein might be described as a

PolyA binding protein whereas closer inspection reveals that the similarity between the candidate protein and true PolyA binding proteins is limited merely to the RNA recognition motif (RRM). Therefore the protein can only be termed with confidence as a RNA binding protein or a protein containing RRM domains. Such an over annotation is not only incorrect but can lead to further errors. Other problems occur when the annotator annotates a single functional domain whilst multiple functional domains are present or the annotator fails to record correctly the actual number of functional domains present. Such errors are common in proteins where multiple domains are prevalent e.g. extracellular receptor signal transduction proteins.

Another problem in using homology information alone is that in many cases the annotation of the database proteins themselves may be incorrect and by utilizing their annotation we simply propagate and proliferate the error. Therefore, it is obviously beneficial if additional evidence for a functional domain can be provided. Initially, the PROSITE database of regular expressions was used to provide diagnostic motifs for protein domains [Bairoch 1993]. However, the use of PROSITE in the *C. elegans* project has now been superseded by the PFAM database. The PFAM database is a collection of hidden Markov models of protein domains [Sonnhammer and Durbin 1997; Bateman *et al.* 1999]. The PFAM database has been constructed to provide a sensitive and accurate automatic method of finding protein domains. Searching this database allows the detection of known domains and accurately records their number and position. As PFAM predicts functional domains, annotations based on PFAM hits can be expressed in a consistent manner. The confidence in PFAM hits allows protein annotation to become semi-automated and easily updated. However a drawback is that

currently PFAM hits are limited to approximately one third of the *C. elegans* predicted protein sequences.

In the cases of proteins where PFAM can establish a putative function a schema for further automated secondary annotation of the protein has been derived. This is a decision tree type process utilizing protein homology to further refine the annotation. For example, the PFAM database contains a model to detect ATP dependent helicases. The PFAM model detects both the DEAD type and DEAH type. The strategy is to take the protein and to search against a small protein database consisting of representative proteins of both DEAD and DEAH type ATP dependent helicases. The premise is that the protein will be of the same class as the protein to which it has the highest score and the protein will adopt the more detailed annotation as long as the score satisfies a previously defined threshold. This strategy relies not only on the fact that proteins with the most similarity in their function will likely share a greater degree of sequence similarity, but that they will also share a more recent common ancestor. Therefore, proteins with the most similar functionality will also share a greater sequence similarity purely due to their closer phylogenetic relationship. In the case of the above ATP helicases where all members fall into one of two subclasses and the family has been well characterized, the threshold is unimportant and can be low. In larger protein families which may not be fully characterised, such as the protein kinases, a higher threshold for the score can be set allowing for a more conservative adoption of the secondary annotation. This strategy has been employed where appropriate for the more abundant PFAM domains found in *C. elegans* e.g. protein kinases, protein phosphatases, RNA recognition motifs and homeobox proteins.

Early in the sequencing project it was realized from the phylogeny of *C. elegans* that little in the way of DNA sequence data from closely related organisms would be readily available and that most of the protein similarities would be between conserved domains predating the divergence of the major animal phyla. Green *et al.* [1993] argued that almost all of these conserved domains, termed ancient conserved regions (ACRs), would already be represented in the protein databases. Figure 1-4 shows the extent of cross-phyla matches to predicted *C. elegans* proteins.

Table 1-3. Cross-phyla matches to *C. elegans* predicted proteins

| **Phyla/Kingdom** | **Database hits** | **%** |
|---|---|---|
| Arthropoda | 4025 | 29.2 |
| Fungi | 4074 | 29.6 |
| Vertebrata | 6230 | 45.3 |
| Prokaryota | 2758 | 20.0 |
| Planta | 3043 | 22.1 |

Matches were calculated as being significant with a BLASTP score =>75 using the Blosum62 matrix. Databases used were Wormpep (release 13 comprising of 13,747 protein sequences), SwissProt (release 35) and Trembl (release 15). Proteins were filtered using the program SEG [Wootton and Federhen 1996] to eliminate matches due to low complexity regions.

In total 55% of *C. elegans* predicted proteins showed cross-phylum/kingdom database matches indicative of ancient conserved domains. The eukaryotic databases in table 1-3 still do not currently represent the full repertoire of ancient conserved domains as 314 *C. elegans* proteins match the prokaryotic protein set only, indicating that searching prokaryotic protein datasets still can be informative. It is also worth noting that although the genome of *Saccharomyces cerevisiae* [Saccharomyces cerevisiae genome sequencing consortium 1997] has been completely derived, only 29.6% of *C. elegans* proteins match fungal proteins

at this threshold. Whether the reason for this is that the divergence between the two organisms is so great that many homologies can no longer be detected using our current methodologies or that functional domains have arisen *de-novo* (or indeed been lost) in the subsequent diverging lineages, it seems clear that comparative sequencing of more closely related species is required to identify all protein coding sequences. To this end, systematic genomic sequencing of the closely related nematode *Caenorhabditis briggsae* has already begun with more that 4 megabases of sequence already derived [M. Marra *et al.* unpublished].

In the absence of more data from closely related species, intra-species protein matches can be extremely useful in genome analysis since many genes are thought to have been derived from the duplication and subsequent divergence of pre-existing genes. Even if the function is unknown, any protein homology supports exon predictions and allows the detection of conserved regions and domains. Intra-genome comparisons are likely to be useful in determining gene structure in the majority of cases where homology is detected. Such intra-family relationships have been instrumental in the determination of gene structure for many of the 7 transmembrane receptor (G-protein coupled receptor) genes in *C. elegans*. These proteins are able to diverge rapidly with few amino acids required to be strongly conserved for their function. In *C. elegans* approximately 650 are thought to exist (A. Bateman personal communication), however few of these show similarity to 7 transmembrane receptors derived from other species. Figure 3 shows the degree of similarity between *C. elegans* proteins. Even with an incomplete set of proteins (Wormpep release 13 containing approximately 75% of *C. elegans* proteins) more than 60% show a BLAST score of 100 or more against one or more other *C. elegans* proteins.

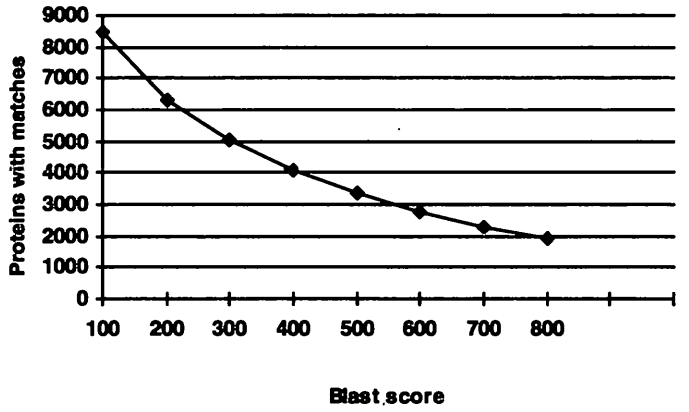



Figure 1-3. Similarity between *C. elegans* proteins. Protein dataset used was Wormpep (release 13) containing 13,747 protein sequences. Matches were calculated using BLAST (using the BLOSUM62 scoring matrix) and filtering protein sequences with SEG.

*Submission to the public sequence databanks*

An important aspect of the *C. elegans* sequencing project is that the sequence data be as readily available to the scientific community as possible. To achieve this, unfinished and unordered sequence contigs of 1Kb or greater from each clone in progress were submitted to the High Throughput Genome (HTG) division of the EMBL/Genbank databanks. The unfinished sequence contigs were updated daily.

The finished *C. elegans* data produced at the Sanger Centre is submitted to the invertebrate division of the EMBL nucleotide databank [Stoesser *et al.* 1998]. ACEDB is capable of outputting sequence data and annotation in the EMBL databank format, which is then submitted to the EMBL databank via e-mail. As the *C. elegans* data is actively curated it is also important that any subsequent changes made to the sequence or its annotation in ACEDB are also propagated to

the public databases. To enable this, a copy of each EMBL file submission is kept and periodically (preferably at intervals of one week) an EMBL format file for all *C. elegans* sequences is made and each file is compared to the most recent EMBL submission using the UNIX diff utility. If a difference is detected between the two files then the new EMBL format file is submitted.

# Organisation of Genomic Information in *C. elegans*

*Introduction*

The organisation of genomic information has been extensively studied in mammalian genomes, most notably that of the human. In the human genome light microscopy reveals chromosomal banding patterns reflecting organisation on the scale of several megabases. Cytogenetics using dye-based chromosomal band technlques has deslgnated a number of recognlsed regions. For, instance the C-band corresponds to the centromere, the G-bands which are detected through a Giesma dye mixture roughly correspond with the regions replicated late, whilst the R-bands define effectively the inverse of the G-bands and replicate early. The T-bands represent a subset of more intensely staining R-bands. The banding pattern of the human genome also correlates with the variety of isochores present. The human genome consists of a mosaic of isochores that are compositionally homogenous regions, which are on average >300kb in length [Bernardi 1993] and the isochores have been divided into families based on their base composition. The GC poor isochores L1 and L2 together represent 60% of the genome, whilst the GC rich isochore H1 represents 10% of the genome and the GC rich isochore H2 makes up 20% of the genome. A further highly GC rich family, H3, makes up a further 5%. It has been shown that the late replicating G bands of the chromosomes are made up almost entirely of GC poor isochores L1 and L2, the R bands (excluding the T band) are made up of approximately equal proportions of GC poor and GC rich isochores. The GC rich isochores are represented mostly by

the H1 isochore family. The T bands consist of GC rich isochores H1 and H2 (with H2 predominating) in conjunction with the very GC rich isochore H3 {Bernardi 1993].

There is a strong correlation between the isochores and gene density, as the increasing GC richness of compartments is associated with increasing gene density. It has been estimated that the G bands represent 50% of the genome and contains only 20% of the genes, whilst the R-bands contain the remaining 80%. The distribution in the types of genes is also believed to be non-random with studies showing that housekeeping[5] genes are found almost exclusively in the R-bands. Conversely, the expression of genes found in the G-bands is more likely to be temporally or spatially restricted to certain tissues of the organism. This observation has important implications in the understanding of genomic organization. It can be proposed that the evolution of multicellular organisms has required the evolution of stringent mechanisms of gene inactivation allowing accurate differentiation of cells and tissues. Thus any increase in cell types and tissue specific transcripts would also increase the pressure to decrease ectopic transcription. Therefore any increase in complexity of an organism can be proposed to be concurrent with not only with an increase in specific mechanisms in transcription but also in its repression. Mechanisms in gene repression include the evolution of nucleosomal chromatin and DNA methylation [Bird and Tweedie 1995]. In addition to these mechanisms it has been suggested tissue specific genes will often be under the transcriptional control of strong promotional and/or

[51]Although the term "housekeeping gene" is commonly used there is no clear or accepted definition. In this context we refer to housekeeping genes as genes which are involved in the processes of cellular metabolism and maintainance and as such are transcribed in most cell types under normal conditions. Such a definition would therefore exclude genes involved in cellular differentiation and genes whose expression is limited to subsets of tissues.

enhancer elements and that the gene sparse environment may aid the tight regulation by reducing crosstalk and noise between adjacent genes.

Genomic organisation in avians has been investigated primarily in the chicken. The chicken karyotype contains 39 chromosomes that are classified into two groups. The six largest chromosomes, termed macrochromosomes, represent 65% of the total genomic DNA [McQueen *et al.* 1998]. The remaining 33 chromosomes are smaller and are termed microchromosomes. Many of the microchromosomes are cytologically indistinguishable due to their small size [Bloom *et al.* 1993]. The microchromosomes and macrochromosomes display very different informational content. The microchromosomes are more GC rich and contain more repetitive sequences than the macrochromosomes [Matzke *et al.* 1990]. However, most striking is the difference between the gene densities of the two compartments. Using the presence of CpG islands [for review see Cross and Bird 1995] and the presence of acetylated histone H4 [for review see Turner 1993] which are both indicators of transcribed regions, it has been estimated that the 65% of the genome represented by the macrochromosomes contains only 25% of the total genes, the majority of the estimated 55,000 genes lying on the microchromosomes [McQueen *et al.* 1998]. This suggests that the gene density on the microchromosomes is one gene per 10KB. The gene rich microchromosomes have also been found to replicate earlier in the S phase which is consistent with other observations that transcriptionally active regions are early replicating [Holmquist 1987]. Interestingly, the majority of genetically mapped loci in the chicken have been found to lie on the macrochromosomes [Burt *et al.* 1995], although it has been proposed that this incongruity is due to biases in the construction of the chicken genome map [McQueen *et al.* 1998].

In *Drosophila melanogaster* genomic organisation is dominated by gene poor and gene rich regions. However, in contrast to the human genome, the gene sparse regions termed heterochromatin are not dispersed but show strong association with the pericentric regions of the chromosomes. Again, as in the human, the gene sparse heterochromatin is replicated later in the S-phase and later than the gene rich euchromatin. The heterochromatin is made up primarily of repetitive DNA sequences and repetitive sequences resembling transposable elements. Evidence also suggests that the heterochromatic regions are transcriptionally inactivated; changes in the boundaries of heterochromatic regions lead to the effect of positional effect variegation in the fruit fly [for review see Elgin 1996]. The euchromatin also displays banding patterns within the polytene chromosomes. The bands are due to the repeated endomitotic replication of the chromosomes which takes place in a number of *Drosophila* tissues, most notably that of the salivary glands. The banding patterns of the *Drosophila* chromosomes have long been used as physical markers allowing the visualization of deletions and rearrangements [Bridges 1935; Kaufman 1939].

*Genomic Organisation in Caenorhabditis elegans*

In the nematode *Caenorhabditis elegans* genomic organisation has been studied primarily through genetic dissection. Each of the five autosomes has been determined to contain two genetically distinct compartments. The central region of each autosome is distinguished by a depressed rate of recombination compared to the flanking arms. This results in high density of loci on the genetic map and has

therefore been termed the "cluster" region. It should be noted that the terminology "cluster" region is potentially confusing and should be distinguished from the term "gene cluster" which refers to a closely associated group of related genes. The sex determining X chromosome in contrast shows less pronounced variation in recombination rate across its length. Previous correlations between the physical map and the genetic map have estimated that each centimorgan within the cluster regions contains approximately 1500KB of DNA. The non-cluster regions display a much higher rate of recombination resulting in 100-300KB of DNA per centimorgan [Barnes *et al.* 1995].

The non-cluster regions of the autosomes have been determined to be enriched in repetitive sequences as compared to the non-cluster regions. Studies of the genomic distribution of repetitive sequences have been carried out [Nacleiro *et al.* 1992, Cangiano and LaVolpe 1993]. These studies suggest that although some repeat elements are distributed evenly across the chromosomes, others such as CeRep3 [Felsenstein and Emmons 1987] and Rc5 [Cangiano and LaVolpe 1993] have been found to be preferentially located within the non-cluster regions. It has been proposed that such repetitive elements, in addition to interstitial telomeric repeats, promote the elevated recombinational rates observed on the autosomal arms [Cangiano and LaVolpe 1993].

Correlations between the genetic and physical maps have also suggested that apart from being recombinationally lowered, the autosomal clusters also contain more genetic loci than the non-cluster regions per megabase of DNA, i.e. non-cluster regions have a lower density of physical genes than the cluster regions. Evidence that the genetic clusters contain a higher physical density of genes has also been supported by EST hybridization data. Barnes *et al.* using a

reference set of 2600 different ESTs indicated that the cluster regions had EST hit rates of 35-52 hits/MB whilst the non-cluster regions had hit rates of 14-30 hits/MB. These observations, however, do not preclude the contrasting interpretation that genes between the two regions may show variability in their ability to display mutant phenotypes or that the genes in the two different regions may show globally differing transcription rates.

Interestingly, the repression of recombination in the gene dense region is counter to the behavior of the chromosomes in other organisms studied. In other organisms studied, the rates of recombination and gene density are positively correlated. It might be argued that this helps maintain an effective rate of allelic assortment and preventing the accumulation of deleterious mutations [Civardi *et al.* 1994; Ikemura and Wada 1991; Mouchiroud *et al.* 1993; Ashburner 1989]. In *C. elegans* the lowered recombination rates across the entire gene dense central clusters result in a tendency for this region to be inherited as a single unit with little recombination between its constituent genes.

Barnes *et al.* [1995] proposed that the establishment of the non-cluster region was due to the selfish expansion of recombination promoting elements in the autosomal arms, whereas the gene dense central regions were sufficiently gene dense not to tolerate an initial invasion/or perpetuation of the recombinational promoting repeats. However, the establishment of the cluster regions is under strong genetic control. Mutations in the *rec-1* gene have been shown to abolish the wild type pattern of genetic exchange causing genetic map distance to correlate with physical distance uniformly across the autosomes [Zetka and Rose, 1995]. *rec-1* mutants show no other obvious phenotype under laboratory conditions. This suggests that for *C. elegans,* the lack of recombination

in the gene dense cluster region has recently provided or continues to provide a selective advantage.

Only a small proportion of genes in the genomic sequence have been studied experimentally (<5%) and almost all functional information acquired for the predicted protein has been inferred computationally, primarily through protein similarity data and the PFAM database of protein domains [Sonnhammer *et al.*1997]. Previous studies have predicted the extent of cross-phyla similarities possessed by *C. elegans* proteins by computational means [Green *et al.* 1993; Sonnhammer 1996]. Such computational methodologies will inevitably lead to an underestimate as sequence divergence may be such that similarity will no longer be detectable using current methodologies. However, homology may subsequently be determined using biochemical analysis, protein structure determination or through the dissection of homologous genetic pathways. The utility of the *C. elegans* as a genetic model organism is also highly dependent on the extent of homology it shares with other organisms especially that of the human. For instance, through the use of suppresser and enhancer screens, the genetic pathway of a human disease gene can be quickly determined if a homologous pathway also exists in *C. elegans* [Ahringer 1997].

Other studies have indicated that the divergence and evolution of *C. elegans* proteins has and probably still continues to be a non-uniform process. Mushegian *et al.* [1998] have shown evidence indicating that approximately 2/3 of genes in *C. elegans* have evolved more rapidly than their homologues within *Drosophila melanogaster.* This intriguing finding suggests that a large subset of *C. elegans* genes have been subjected to very different evolutionary pressures since their divergence from a common ancestor with the fruit fly. It is likely that these

unequal evolutionary rates with *C. elegans* has provided the ambiguity in the assignment of the evolutionary branching between arthropods, nematodes and vertebrates [Sidow and Thomas 1994; Nielsen 1995].

*Methodology*

Data was taken from *C. elegans* ACEDB release WS6 [Durbin and Thierry-Mieg 1991-]. Marey maps for each of the chromosomes were constructed using genetically mapped loci which had been correlated to either a precise open reading frame or to a sequenced clone derived from the physical map. GFF format [Durbin *et al.* 1997-] files for each chromosome were derived using GIFACE [Durbin and Thierry-Mieg 1991-]. The GFF format of chromosomal features allows the easy computational interrogation and manipulation of chromosomal data using PERL and the GFF PERL module (T. Hubbard, unpublished). All protein homology searches were carried out using BLASTP version 2 [Gish unpublished 1991-1997] and utilized SEG for low complexity filtering [Wootton and Federhen 1996]. Similarities were only considered where the probability of the same match occurring randomly was $10^{-3}$ or less.

The non-nematoda protein database used in table 2-1 was derived from SwissProt release 36 using SRS [Etzhold *et al.* 1996]. The *C. elegans* EST dataset used contained 72,451 ESTs derived from Waterston *et al.* [1992], Adams *et al.* [1991] and Kohara [1996]. Estimates of the genetic cluster boundaries where taken from Barnes *et al.* [1995].

Repeat family elements were mapped to genomic sequence using the HMMER Hidden Markov modelling package [Eddy 1995-]. More detailed

information of *C. elegans* repeat families and their nomenclature can be obtained from http://www.sanger.ac.uk/Projects/C_elegans/repeats/.

In determining cross-species protein similarities with *C. elegans*, yeast proteins were derived from the ORF set maintained in the Saccharomyces Genome Database [http://genome-www.stanford.edu/Saccharomyces/], human proteins were derived from SwissProt version 36 (RL41_HUMAN was excluded from this study due to its short size of 25 amino acids) and *E. coli* proteins were derived from the set maintained at the NCBI Entrez genome division [http://www.ncbi.nlm.nih.gov/Entrez/Genome/org.html].

Putative 7 transmembrane receptor proteins (G protein coupled receptors) were detected using a Hidden Markov Model (HMM) trained using known *C. elegans* 7 transmembrane receptor proteins (A. Bateman, personal communication).

Putative orthologs were defined as being reciprocally the most similar pairs. For example, each *C. elegans* protein was searched against the canonical yeast protein dataset. A BLASTP (ver. 2) threshold of $E<=1X10^{-3}$ was employed and low complexity filtering of the sequences was achieved using SEG. The highest scoring yeast protein was then searched against the *C. elegans* protein dataset using the same thresholds and low complexity filter. If the most similar *C. elegans* protein to the yeast protein was the same *C. elegans* protein which detected the yeast protein initially then putative orthology between the two proteins was established.

The *C. elegans* dataset used consisted of 95.5Mb of DNA sequence. At the time of writing not all the sequence had been completely confirmed to be within the

desired accuracy of 1 error in 10Kb of DNA. Sequences still in progress at the time of the study were included, although the gene predictions from these sequences were based only on those provided by GENEFINDER [Green and Hillier, unpublished] and no attempt to manually appraise these genes was made. A number of sequence gaps also remain (approximately 100). However, the extent of the missing sequence is thought to be small since well in excess of 99% of the EST sequences match the *C. elegans* sequence already determined. This figure is also confirmed by the sequencing of random clones from a whole genome library [The *C. elegans* Sequencing Consortium 1998].

*Results*

A summary of the genes found in each of the chromosomes and their genetically defined compartments is shown in table 2-1. Marey maps for the six chromosomes showing the relationship between recombinational rate and physical DNA distance are shown in figure 2-1. The boundaries of the autsomal cluster regions are also shown in figure 2.1. The cluster boundaries defined by Barnes *et al* 1995 have been adopted. It should be noted that the boundaries positions are based on genetic mapping data which remains incomplete e.g. lack of mapping data means that the cluster regions could be expanded on the right arm of chromosome II and the left arm of chromosome V.

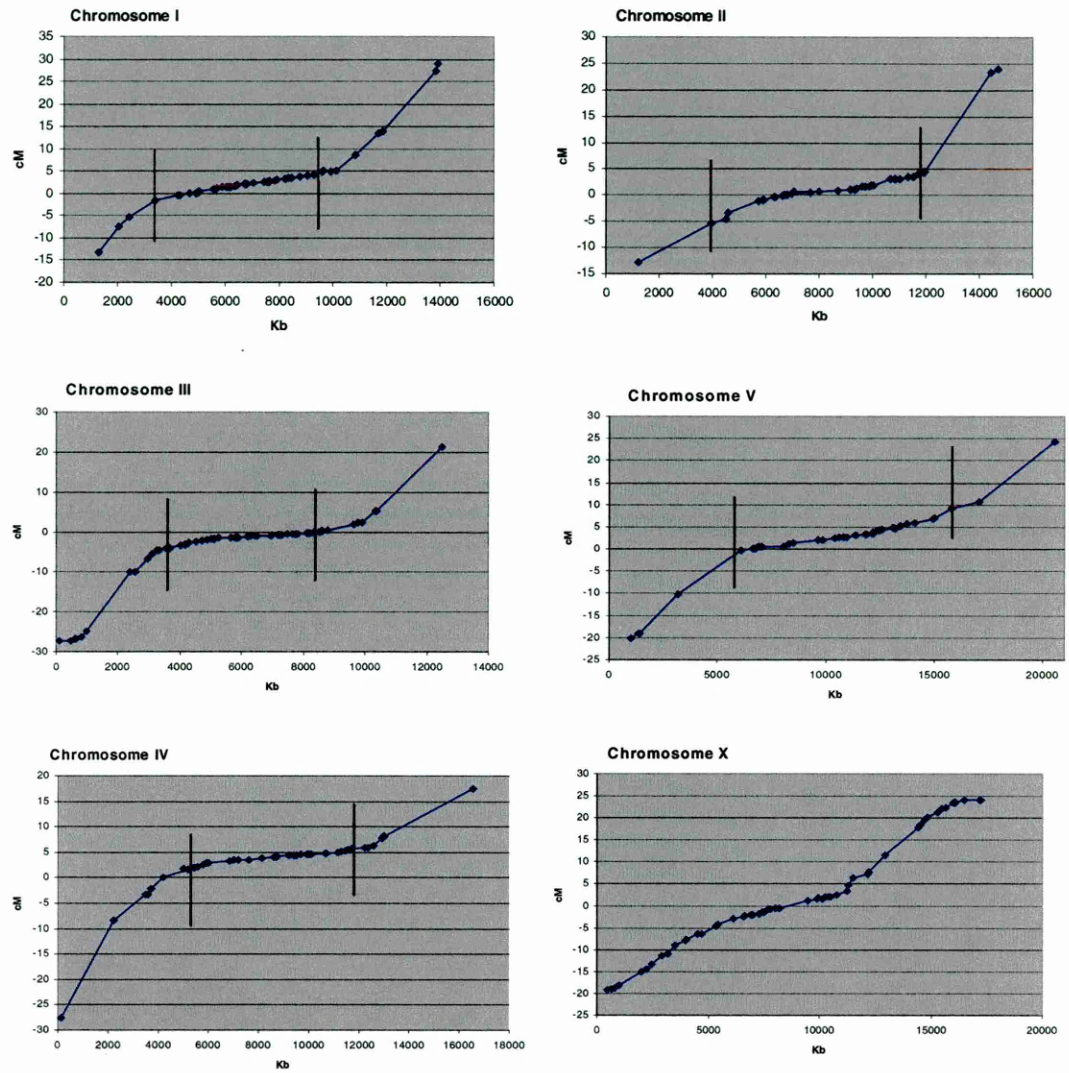



Figure 2-1: Marey maps for the six *C. elegans* chromosomes. The extent of the autosomal central cluster regions as defined by Barnes *et al.* 1995 is marked.

Table 2-1: Gene distribution in the *C. elegans* genome

| Chromosome[1] | | Size (Mb) | Protein Genes | Density (Kb gene$^{-1}$) | TRNA genes[2] | % Coding | % EST match[3] | % Database Match[3,4] |
|---|---|---|---|---|---|---|---|---|
| I | L | 3.29 | 649 | 5.06 | 7(2) | 21.59 | 57.0 | 53.9 |
| | C | 5.59 | 1171 | 4.77 | 34(4) | 31.65 | 52.9 | 52.1 |
| | R | 4.98 | 983 | 5.06 | 33(2) | 25.00 | 43.4 | 40.8 |
| II | L | 3.83 | 1049 | 3.65 | 29(13) | 29.00 | 22.7 | 26.9 |
| | C | 7.93 | 1719 | 4.61 | 38(6) | 29.68 | 49.7 | 49.8 |
| | R | 2.96 | 491 | 6.03 | 16(5) | 19.89 | 43.5 | 39.9 |
| III | L | 3.30 | 612 | 5.4 | 31(14) | 20.60 | 44.2 | 42.1 |
| | C | 4.98 | 1100 | 4.52 | 42(0) | 32.21 | 53.5 | 53.5 |
| | R | 4.49 | 796 | 5.66 | 21(3) | 23.91 | 53.1 | 50.2 |
| IV | L | 5.44 | 1050 | 5.17 | 38(16) | 20.87 | 39.9 | 39.7 |
| | C | 6.51 | 1422 | 4.58 | 20(3) | 29.69 | 45.7 | 50.3 |
| | R | 4.19 | 622 | 6.73 | 26(2) | 16.5 | 36.6 | 40.7 |
| V | L | 6.19 | 1491 | 4.15 | 17(4) | 27.00 | 22.0 | 33.0 |
| | C | 6.84 | 1573 | 4.34 | 37(0) | 29.40 | 32.2 | 43.8 |
| | R | 7.79 | 1782 | 4.36 | 152(94) | 25.50 | 19.5 | 28.8 |
| X | | 17.22 | 2631 | 6.54 | 362(33) | 19.8 | 40.9 | 43.34 |
| Total | | 95.53 | 19141[5] | 4.99 | 877(198) | 25.1 | 40.3 | 43.2 |

[1]Autosomes are divided into the genetically defined compartments, left arm (L), central cluster region (C) and right arm (R).
[2]Parenthesis denote the number low scoring predictions thought to be pseudogenes [Lowe and Eddy 1997].
[3]The % of genes with EST and Database matches was determined only from manually appraised genes.
[4]Matches to non nematoda proteins determined using BLASTP version 2 where P<=0.001.
[5] This total represents the total number of protein coding elements, including those from contigs not as yet mapped to a chromosome and alternative transcripts.

The distribution of genes in the *C. elegans* genome [table 2-1] shows that for most autosomes the gene density within the recombinationally suppressed central cluster regions is higher than in the flanking arms. The notable exceptions being chromosome V which shows a relatively high gene density across its entire length and the left arm of chromosome II which shows the highest gene density. However, it should be noted that the gene densities and percentage coding do not correlate strongly e.g. although the left arm of chromosome II has the highest gene density it does not have the highest percentage of predicted protein-coding bases. This suggests that either some regions have a proliferation of smaller genes or that the gene finding methodologies have been less effective in certain regions.

tRNA gene distribution also shows regional variation. The X chromosome shows the highest density of tRNA genes with 41% of the total predicted tRNA genes for *C. elegans* found on this chromosome. The right arm of V has 152 predicted tRNA genes, however, 61% of these have poor secondary structures consistent with being non-functional pseudogenes. The proportion of predicted tRNA genes that are thought to be pseudogenes also varies significantly between regions. The left arms of II, III, IV and the right arm of V show pseudogene rates in excess of 44%, whilst for the remainder of the genome the tRNA pseudogene rate is below 10%.

The proportion of genes which are represented in the EST dataset also show regional specificity. The left arm of I shows the highest EST hit rate at 57%, whilst the lowest EST rate is found on the right arm of chromosome V at 19.8%. The variation of database matches with the conceptual protein translations varies between 26.9 and 53.9%. The proportion of genes with a database hit also shows a positive correlation with proportion of genes which possess an EST. The correlation coefficient between EST hit rate and database hit in the 16 chromosomal compartments is 0.93, which exceeds 99% confidence limits.

*Gene Structure*

The median intron and exon sizes are 65bp (n=97,879) and 150bp (n=117,231) respectfully. Gene structure was not found to be uniform across the chromosomes. Figures 2-2a,b and 2-3a,b show the variation in intron and exon median sizes across the six chromosomes. The median intron size shows an increase in many of the autosomal arms compared to their corresponding central

cluster regions. Exceptions to this are the left arms of II and V, which have also been associated with high gene density (table 2-1). However, the right arm of V has a similar gene density to its left arm but does show an increase in median intron size. The X chromosome shows an increase in median intron size towards both extremities and also at its centre. The X chromosome also shows an overall increase in median intron size compared to the autosomes.

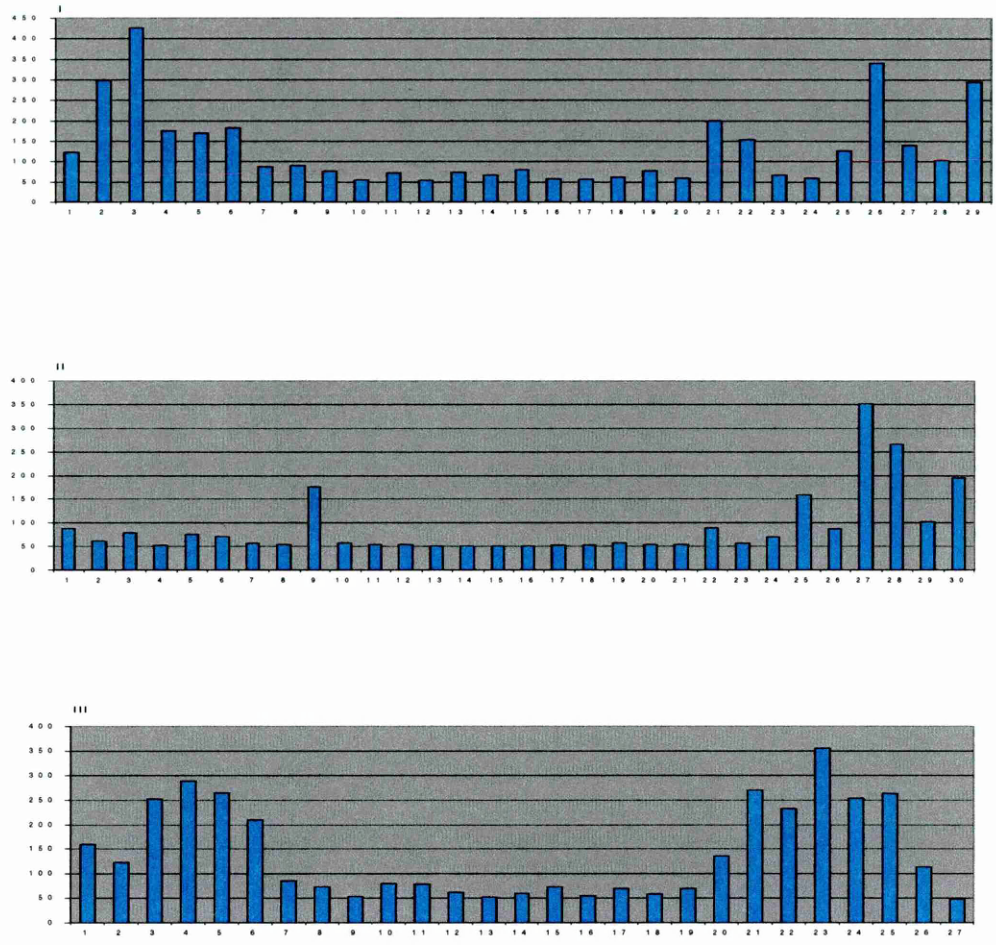



Figure 2-2a: Median intron sizes across chromosomes I, II and III. The Y-axis indicates intron size in base pairs. Each X-axis interval represents 500kb of genomic sequence

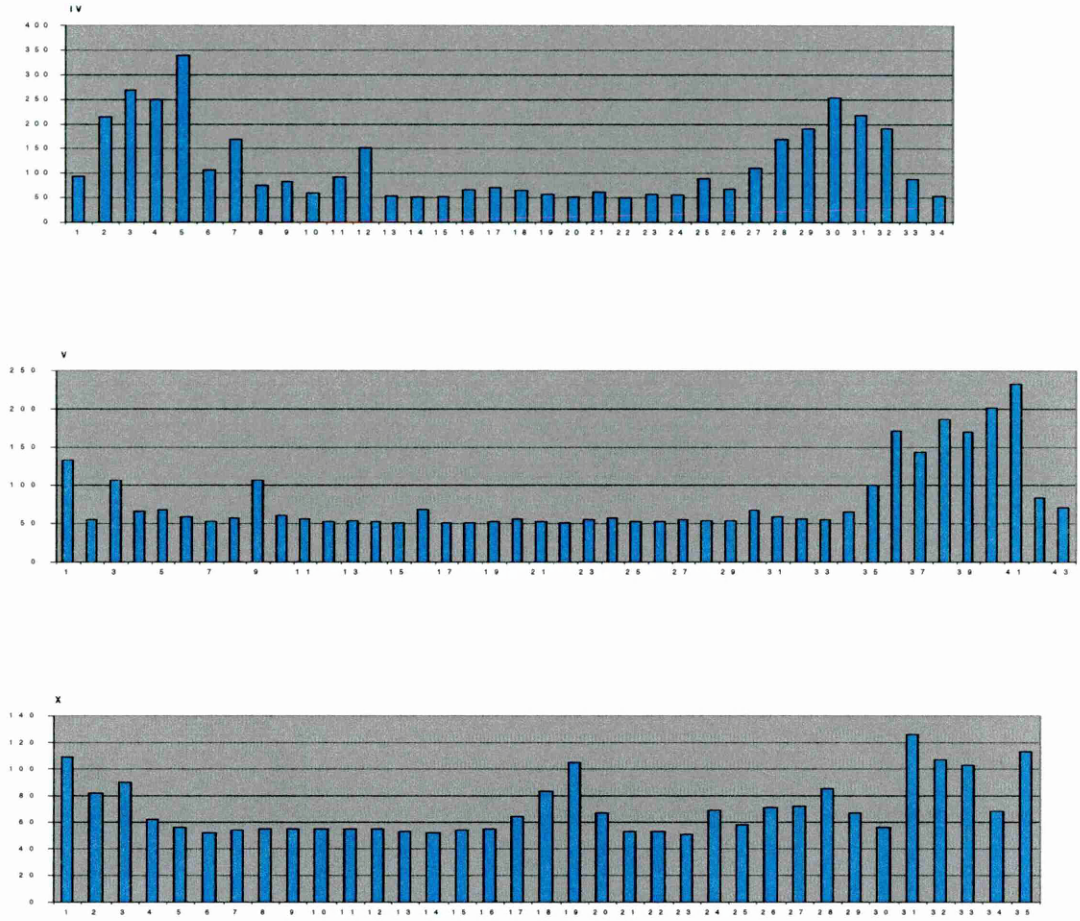



Figure 2-2b: Median intron sizes across chromosomes IV, V and X. The Y-axis indicates intron size in base pairs. Each X-axis interval represents 500kb of genomic sequence.

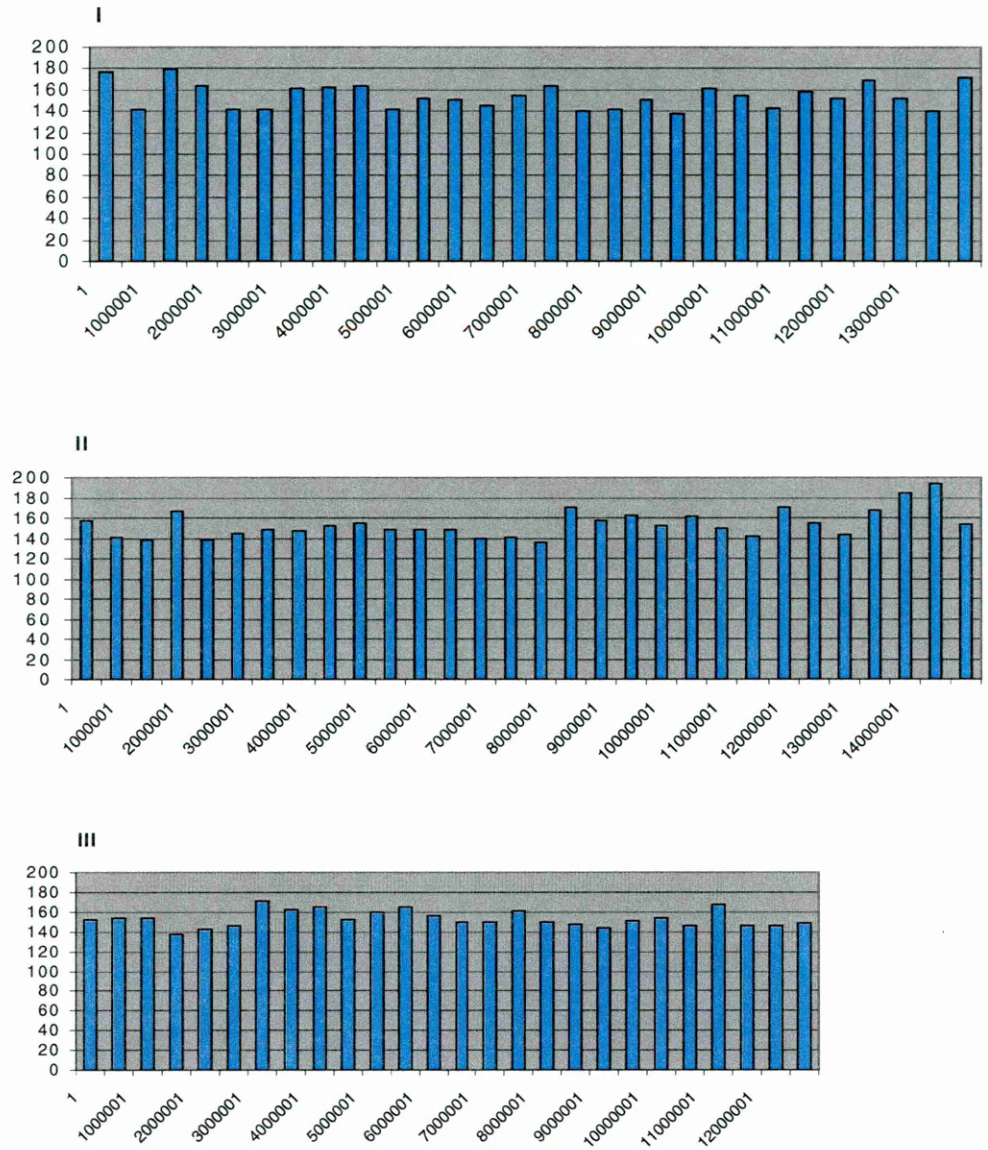



Figure 2-3a: Median exon sizes across chromosomes I, II and III. Y-axis shows exon size in base pairs. Each X-axis interval represents 500kb of genomic sequence.

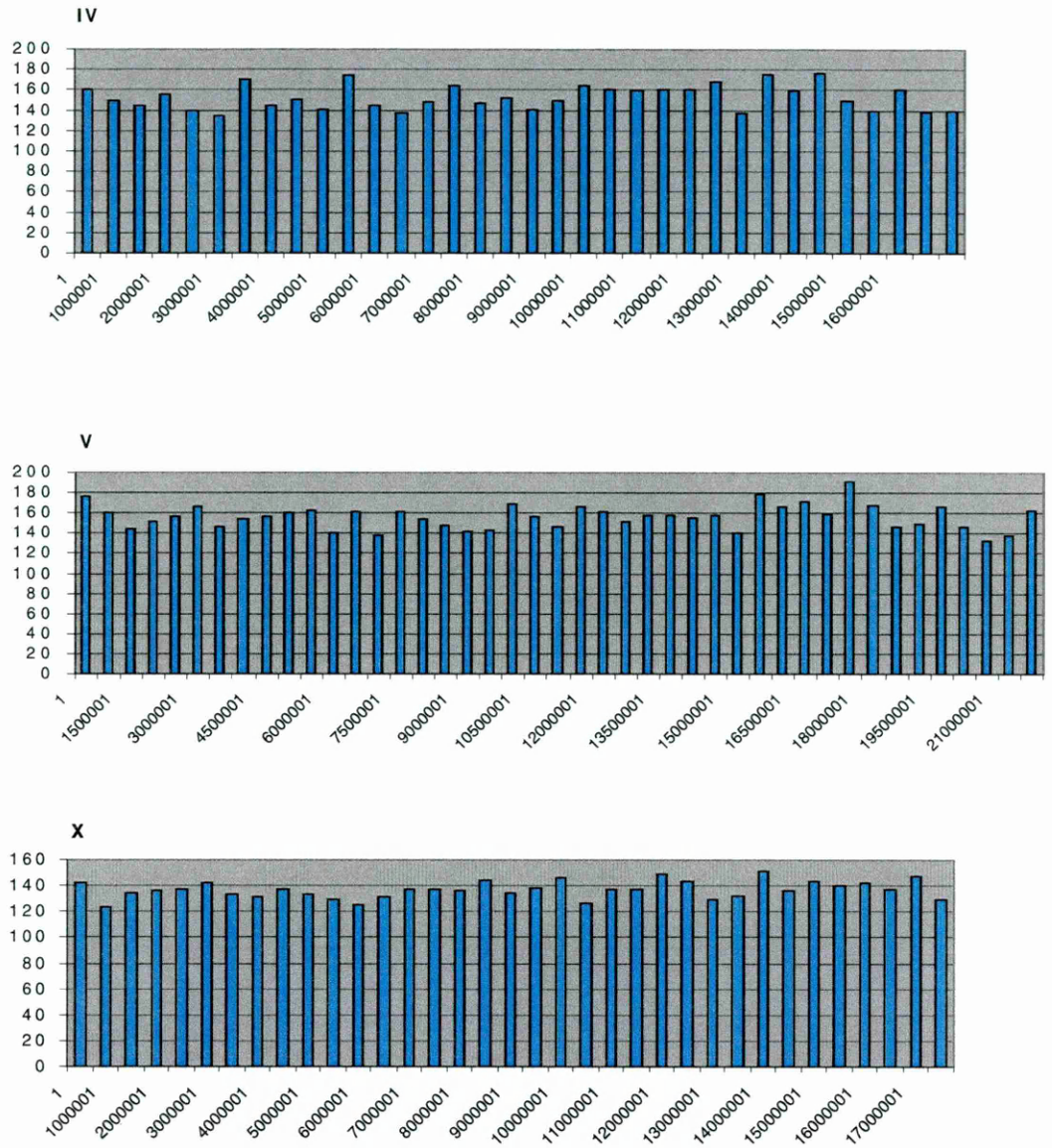



Figure 2-3b: Median exon sizes across chromosomes IV, V and X. Y-axis shows median size in base pairs. Each X-axis interval represents 500kb of genomic sequence.

The distribution in exon size does not vary greatly across the chromosomal compartments, although the median exon length within the X chromosome was found to be gene a little lower on the autosomes. On the X chromosome the median exon size was found to be (136bp (n=19,014) compared to 150bp(n=98,217)).

*Repeat Distributions*

The chromosomal distribution of repetitive elements has also been studied. In this case both simple repeat elements such as tandemly repeated DNA sequence and inverted repeats, and a number of families of repeat elements were identified [R. Durbin unpublished; Eddy and Lewis unpublished]. The distribution of repeat elements is shown in table 2-2.

Inverted and tandem repeats both show over representation in intronic sequence. They also both show a strong preference for highly recombinagenic regions i.e. they show a preference for the autosomal arms and are under-represented on the autosomal central cluster regions as well as the X chromosome. A small proportion of inverted and tandem repeat elements overlap with exonic sequence. The presence of tandem repeats in exon sequence is not unexpected as *bona-fide* protein coding elements can contain tandemly repetitive sequence, therefore the presence of exonic tandem repeats is not necessarily indicative of an error in gene prediction. For example, the approximately 170 cuticle collagen genes present in the genome often contain strongly conserved tandem repeats coding for the poly(Gly-X-Y) domain.

Table 2-2. Repeat element distribution

| Repeat Type | MB covered | % on X chromosome | % on autosome arms | % on autosome clusters | % intronic | % intergenic |
|---|---|---|---|---|---|---|
| (Non-Coding) | 71.53 | 19.3 | 49.7 | 31.0 | 35.1 | 64.9 |
| Tandem | 3.16 | 6.7 | 80.2 | 13.0 | 43.9 | 51.7 |
| Inverted | 4.10 | 10.4 | 72.0 | 17.6 | 43.3 | 51.9 |
| CeRep10 | 0.352 | 1.5 | 48.0 | 50.5 | 55.1 | 43.3 |
| CeRep11 | 0.070 | 0 | 59.9 | 40.0 | 58.1 | 39.8 |
| CeRep12 | 0.246 | 8.3 | 62.1 | 29.6 | 38.3 | 61.1 |
| CeRep13 | 0.038 | 21.6 | 43.3 | 35.0 | 33.9 | 65.3 |
| CeRep14 | 0.209 | 1.0 | 81.5 | 17.5 | 48.5 | 50.4 |
| CeRep15 | 0.060 | 9.9 | 56.8 | 33.3 | 51.3 | 46.6 |
| CeRep17 | 0.094 | 7.2 | 84.7 | 8.2 | 39.0 | 59.1 |
| CeRep18 | 0.055 | 26.4 | 42.0 | 31.6 | 32.5 | 66.4 |
| CeRep19 | 0.560 | 2.9 | 93.2 | 3.9 | 48.2 | 49.4 |
| CeRep20 | 0.063 | 1.0 | 68.3 | 30.7 | 60.5 | 36.7 |
| CeRep21 | 0.040 | 17.9 | 46.2 | 35.9 | 36.8 | 60.9 |
| CeRep22 | 0.011 | 4.5 | 50.2 | 45.3 | 43.6 | 51.4 |
| CeRep23 | 0.179 | 18.6 | 56.7 | 24.8 | 40.0 | 58.0 |
| CeRep24 | 0.234 | 10.7 | 83.7 | 5.6 | 47.4 | 51.2 |
| CeRep25 | 0.014 | 0 | 99.6 | 0.4 | 76.6 | 22.7 |
| CeRep26 | 0.242 | 4.3 | 89.6 | 6.1 | 49.1 | 47.1 |
| CeRep27 | 0.113 | 26.0 | 48.4 | 25.6 | 7.9 | 91.5 |
| CeRep28 | 0.048 | 18.8 | 58.6 | 22.5 | 40.9 | 58.7 |
| CeRep29 | 0.122 | 11.3 | 72.0 | 16.8 | 33.8 | 65.3 |
| CeRep30 | 0.026 | 14.7 | 50.5 | 34.8 | 31.5 | 64.5 |
| CeRep31 | 0.032 | 0 | 21.4 | 78.6 | 23.0 | 76.1 |
| CeRep32 | 0.130 | 7.0 | 86.0 | 7.0 | 39.1 | 60.1 |
| CeRep33 | 0.019 | 30.9 | 38.8 | 30.3 | 29.5 | 68.7 |
| CeRep34 | 0.217 | 40.6 | 39.7 | 19.8 | 37.3 | 61.3 |
| CeRep35 | 0.258 | 6.4 | 86.3 | 7.3 | 33.4 | 64.2 |
| CeRep36 | 0.179 | 11.5 | 72.6 | 15.9 | 35.6 | 62.9 |
| CeRep40 | 0.138 | 8.3 | 78.2 | 13.5 | 30.9 | 65.7 |
| CeRep41 | 0.030 | 20.3 | 54.1 | 25.6 | 22.4 | 76.1 |
| CeRep42 | 0.066 | 54.2 | 34.8 | 11.0 | 32.9 | 66.3 |
| CeRep43 | 0.457 | 35.3 | 51.5 | 13.2 | 37.8 | 61.0 |

Repeat family elements also show regional specificity. CeRep26, which corresponds to the telomeric hexamer repeat TTAGGC, is found outside the telomeric regions but shows strong preference for the autosomal arms. Other repeats also show a strong autosomal arm preference such as CeRep14,

CeRep17, CeRep19, CeRep25, CeRep32 and CeRep35. Conversely, CeRep31 shows a strong propensity for autosomal cluster regions. CeRep10 and CeRep11 both are under-represented on the X chromosome [this observation was first made by Richard Durbin].

Many of the repeat families show positive bias towards intronic sequences, especially CeRep25. Other repeats such as CeRep31 and CeRep41 show a bias toward intergenic sequences.

*Patterns of protein homology across the chromosomes.*

The degree of protein conservation between *C. elegans*, *Saccharomyces cerevisiae*, *E.Coli* and human is represented in figure 2-4. All proteins sets except that of human are derived from complete (or near complete) genomes. Unsurprisingly, *C. elegans* showed most extensive similarity with its closest relative represented in the figure, the human. 74% of the human proteins represented in the database show similarity to *C. elegans* proteins. This indicates that the more complex and gene rich human genome has extensively used and re-used functional domains that were present in the most recent ancestor shared between these two organisms. However, it should be noted that the human protein set studied here is incomplete and will likely be biased in its composition towards proteins existing in multiple phyla. The distribution of *C. elegans* genes with putative orthologs from *Saccharomyces cerevisiae* and human is shown in figures 2-5a and 2-5b. Putative orthologs were determined as being defined as reciprocally the most similar pairs.

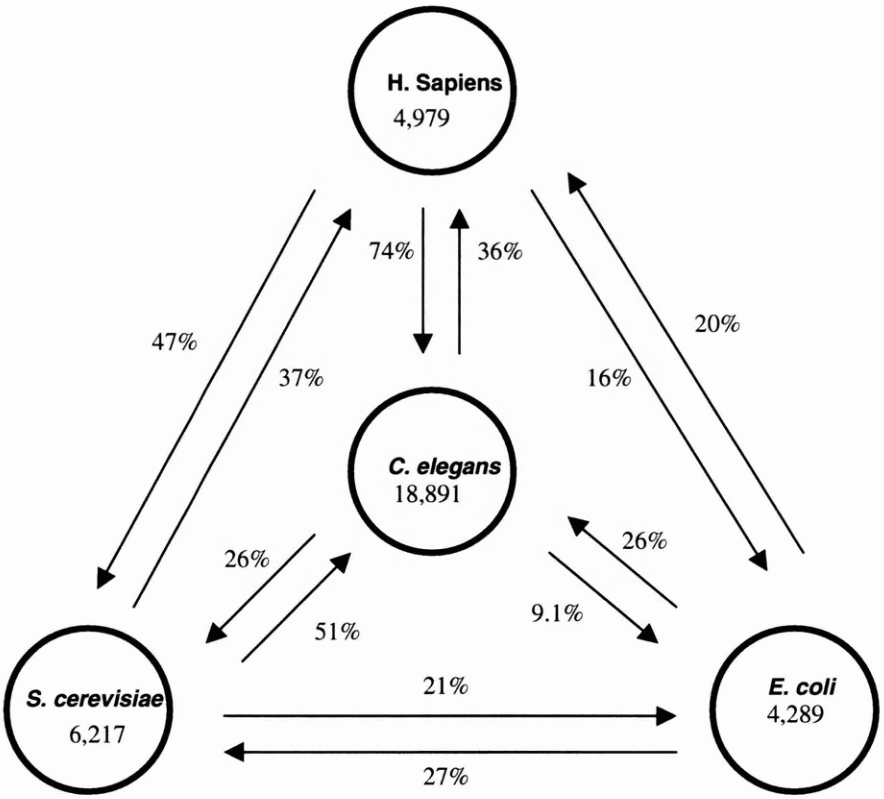



Figure 2-4: Percentages of similar proteins between *C. elegans*, Human, *E. Coli* and *S. cerevisiae*. This figure is an adaptation of a similar figure from Sonnhammer [1996].

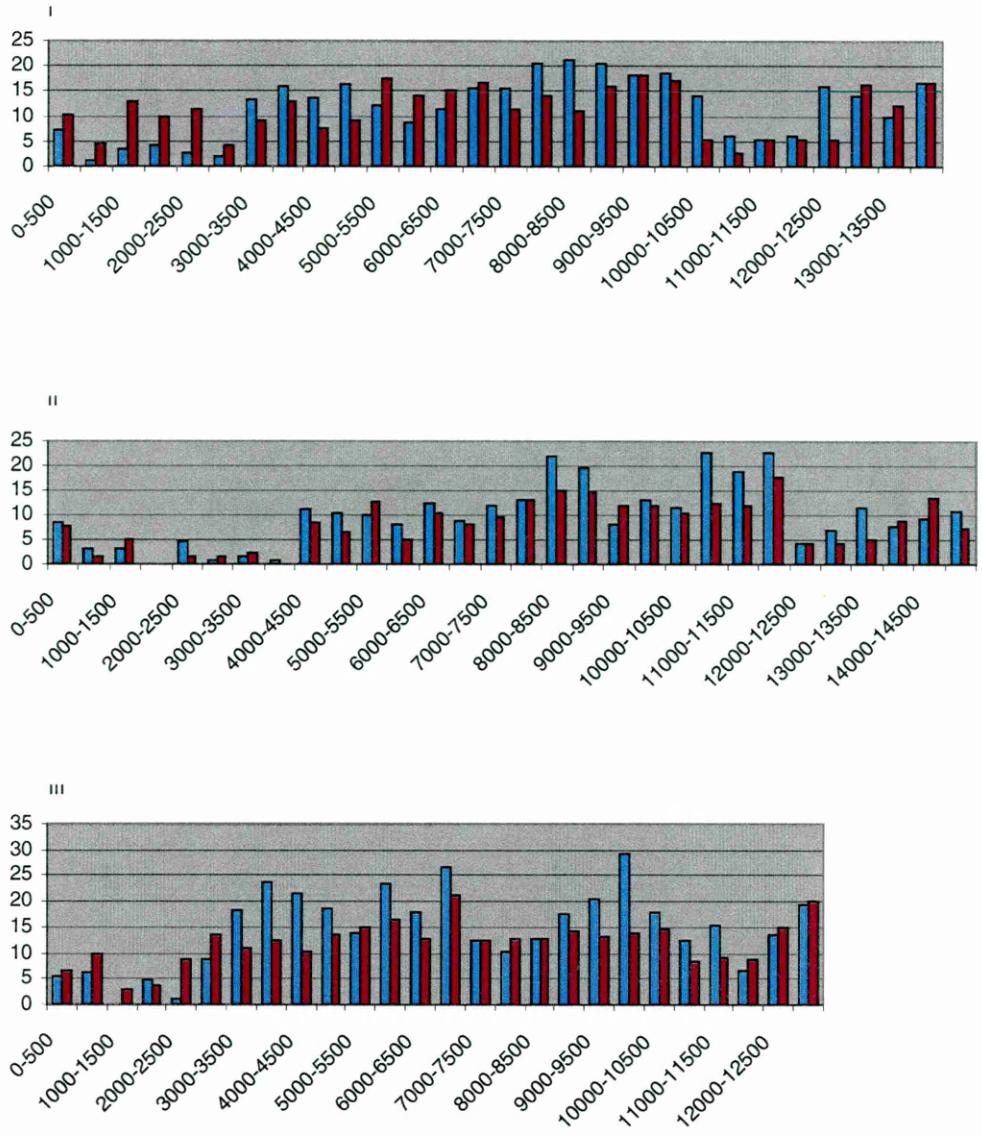



Figure 2-5a: Putative orthologues of *C. elegans* with yeast and human. The Y-axis shows the percentage of genes in each 500kb window possessing a putative homologue. Blue represents orthologues with yeast and red with human. Genetic cluster boundaries determined by Barnes *et al.* [1995] are at: I 3.383Mb and 9.585Mb; II 3.934Mb and 11.906Mb; III 3.51Mb and 8.513Mb.

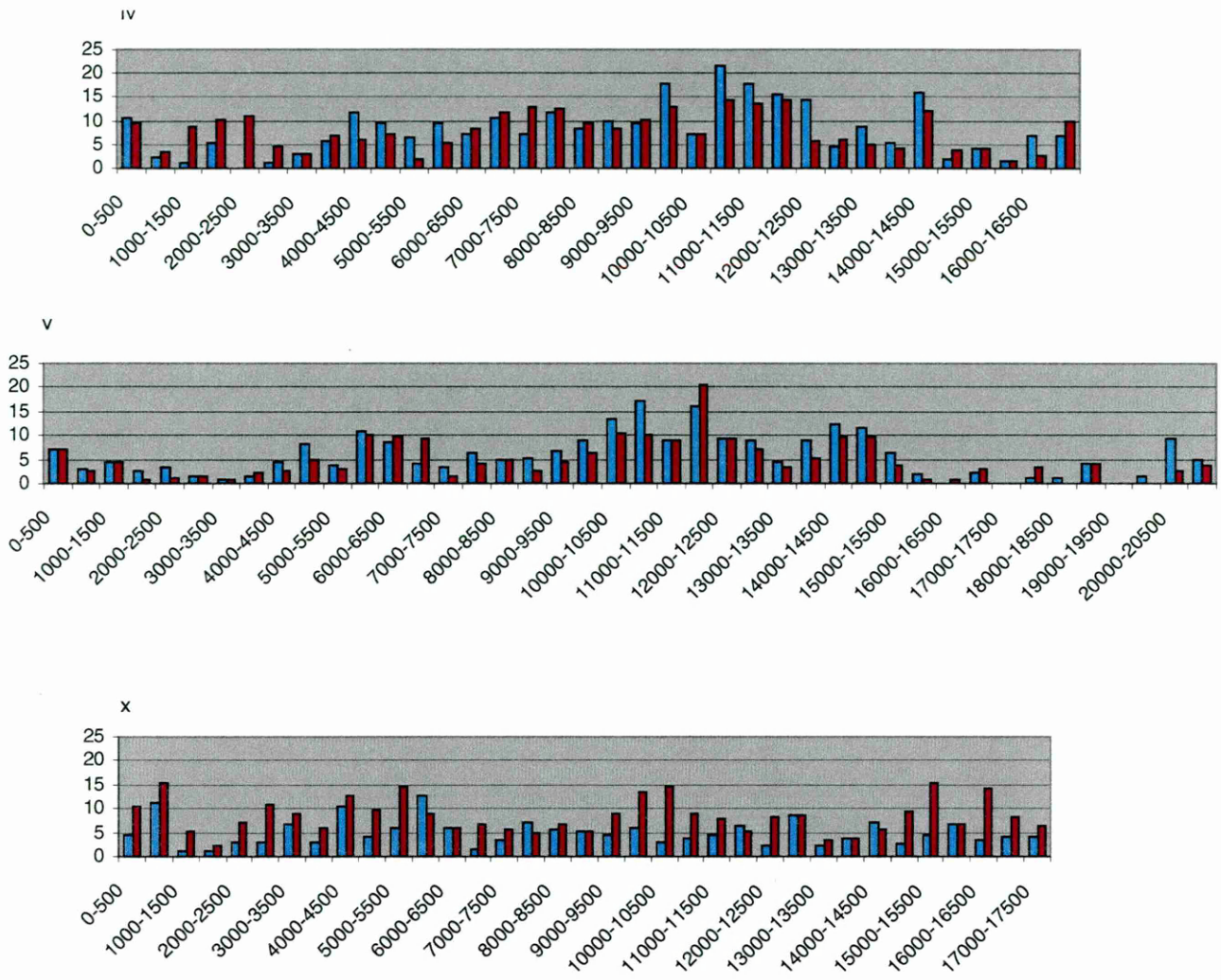



Figure 2-5b: Putative orthologues of *C. elegans* with yeast and human. The Y-axis shows the percentage of genes in each 500kb window possessing a putative homologue. Blue represents orthologues with yeast and red with human. Genetic cluster boundaries determined by Barnes *et al.* [1995] are at: IV 5.543Mb and 12.555Mb; V 6.522Mb and 13.396Mb.

Previous studies predicted that the genetic cluster regions would have a higher gene density than the autosomal arms [Barnes *et al.* 1995]. Such estimates have been based on the distribution of genetic loci as well as cDNA hybridization studies. The efficacy of such an approach to estimate gene density is dependent on the mutational detectability of genes being constant across the chromosomes and also dependent upon overall transcription rates being uniform across the chromosomes.

In order to determine whether differences in the genetic accessability of genes exist across the chromosomes, genetically defined loci were identifed for each chromosomal compartment. Lethal (*let*) genes were excluded from this particular assay, as they have been predominately derived from genetically balanced regions. The results are summarized in table 2-3. The percentage of genes hit in classical genetic studies was found to range from 8.28% in the central cluster region of chromosome I to 1.14% on the left arm of chromosome II. On all autosomes except chromosome III the cluster region was found to be denser in genetic loci than the flanking arms, although in most cases this is not currently statistically significant.

Table 2-3 indicates that different autosomal compartments show different degrees to which they have been amenable to classical genetic analysis. Such differences could be accounted for if there has been an amplification of genes in the arms so that they are more likely to possess semi-redundant or overlapping functions and therefore less likely to be detected in mutagenesis screens.

Table 2-3:Correlation between mutationally defined loci and predicted genes.

| Chromosome | | Boundary | Map Position | # of loci | Predicted Genes | % of genes hit[1] | # of genes in clusters[2] | % of genes in clusters[2] |
|---|---|---|---|---|---|---|---|---|
| I | L | *Unc-73* | -1.899 | 29 | 649 | 4.47+/- 1.39 | 38 | 8.50 |
| | C | | | 97 | 1171 | 8.28 +/- 1.37 | 82 | 7.00 |
| | R | *Lin-11* | 4.837 | 44 | 983 | 4.48 +/- 1.12 | 182 | 18.51 |
| II | L | *Lin-31* | -5.2 | 12 | 1049 | 1.14 +/- 0.57 | 373 | 35.55 |
| | C | | | 106 | 1719 | 6.17 +/- 0.99 | 227 | 13.21 |
| | R | *Lin-29* | 4.257 | 24 | 491 | 4.89 +/- 1.63 | 94 | 19.14 |
| III | L | *Dpy-27* | -4.406 | 26 | 612 | 4.25 +/-1.47 | 68 | 11.11 |
| | C | | | 86 | 1100 | 7.82 +/- 1.36 | 81 | 7.36 |
| | R | *Glp-1* | 0.145 | 65 | 796 | 8.17 +/- 1.63 | 81 | 10.18 |
| IV | L | *Skn-1* | 2.143 | 48 | 1050 | 4.57 +/- 1.14 | 202 | 19.24 |
| | C | | | 80 | 1422 | 5.63 +/- 1.05 | 215 | 15.12 |
| | R | *Unc-31* | 6.282 | 15 | 622 | 2.41 +/- 1.13 | 136 | 21.86 |
| V | L | *StP192* | -0.014 | 22 | 1491 | 1.46 +/- 1.29 | 549 | 36.82 |
| | C | | | 74 | 1573 | 4.70 +/- 0.95 | 372 | 23.64 |
| | R | *LwP6* | 5.51 | 36 | 1782 | 2.02 +/- 0.56 | 546 | 30.63 |
| X | | | | 150 | 2631 | 5.70 +/- 0.76 | 265 | 10.07 |

[1]Error bars reflect 95% probability within a binomial distribution.
[2] Groups of closely related genes. See section 5 "Gene clusters in *C. elegans*".

Table 2-3 also shows an estimate of number of genes from each compartment derived from gene clusters. Gene clusters being defined as groups of closely related genes [as discussed in Part 5]. There is a negative correlation coefficient of -0.794 between the percentage of genes within clusters and the percentage of genes mutationally defined (which is significant at the 99% confidence level).

In addition it has been shown (this thesis) that mutationally defined loci in *C. elegans* are more likely to be represented within the EST dataset than the overall EST coverage would suggest. The correlation coefficient between the EST representation of genes in each compartment (table 2-1) and the percentage of genes mutationally defined is 0.84, which exceeds 99% confidence limits.

The number of genetically defined loci is also positively correlated with the proportion of genes possessing a database hit (table 2-1). The correlation coefficient of this relationship is 0.85, which again exceeds 99% confidence limits.

Therefore we can invoke at least three factors influencing the genetic amenability in *C. elegans* i.e. EST representation, redundancy and conservation. It should be noted that these correlations could be strongly influenced by a relatively small number of gene families. For example, over 650 7 transmembrane chemoreceptor genes are thought to exist in the genome [Robertson 1998; *C. elegans* Sequencing Consortium, 1998]. These genes are almost absent in the EST datasets, but are commonly found in gene clusters [Troemel *et al* 1995; this thesis] and their function in roles such as perception of specific oderants would mean that no obvious phenotype would be observed in the vast majority of mutagenic screens carried out in *C. elegans* genetics. From the distribution of the 7 transmembrane receptors [figures 2-5a and 2-5b] it can be seen that this class of gene is highly represented on chromosome V as well as the left arm of chromosome II. These two regions show the lowest density of mutationally defined loci. However, if only approximately 650 7 transmembrane receptors exist in the entire genome then this single class alone cannot be responsible for the decrease in the density of mutationally defined loci on chromosome V and the left arm of II. The relatively high numbers of 7 transmembrane receptors are more likely to be indicative of the increased overall redundancy in these regions due to the increased presence of gene clusters (see section 5 "Gene clusters in *C. elegans*").

It is unclear, however, how representative visible loci are in such an appraisal. Visible phenotypes in *C. elegans* are mostly due to muscle, neuronal

and cuticle defects and therefore exclude many genes involved in essential cellular processes. More detailed genetic studies have also been carried out in essential genes, albeit in smaller segments of the genome. Three studies of essential gene content have been carried out within recombinationally balanced regions on I (left) [Johnsen *et al.* submitted]; III (left) [Stewart *et al.* 1998] and V (left) [Johnsen and Baillie 1991]. In the light of genomic sequence these studies can now be examined knowing the full physical content of these regions. Table 2-4 shows the essential loci content in each of these regions as well as the estimated minimal essential gene content as predicted from the Poisson distribution estimates from each study. As each balanced region also traverses an arm/cluster boundary, the number of loci from each of these compartments is also described. In addition, the proportion of genes which are part of a gene cluster as well as the proportion of genes being putative yeast orthologs (defined as being reciprocally the most similar pairs in both species protein datasets) are shown. The results from this analysis show that different regions of the genome possess different densities of essential loci. On the left arm of chromosome I, 33% of genes are predicted to confer an essential function whilst in total less than 5% on the genes on V (left) are essential. The essential gene rich region on I(left) is characterised by having a relatively high proportion of putative yeast homologs and a low proportion of genes within gene clusters. The region with the smallest proportion of essential genes is characterised by having a relatively low proportion of putative yeast homologs and a high proportion of genes within gene clusters. The essential gene content also differs between the genetically defined arm and cluster regions. On chromosome I the difference is slight being only around 10%. In the other two regions from chromosomes III and V, increases of approximately

100% in the expected number of essential loci are observed between the arm and cluster regions.

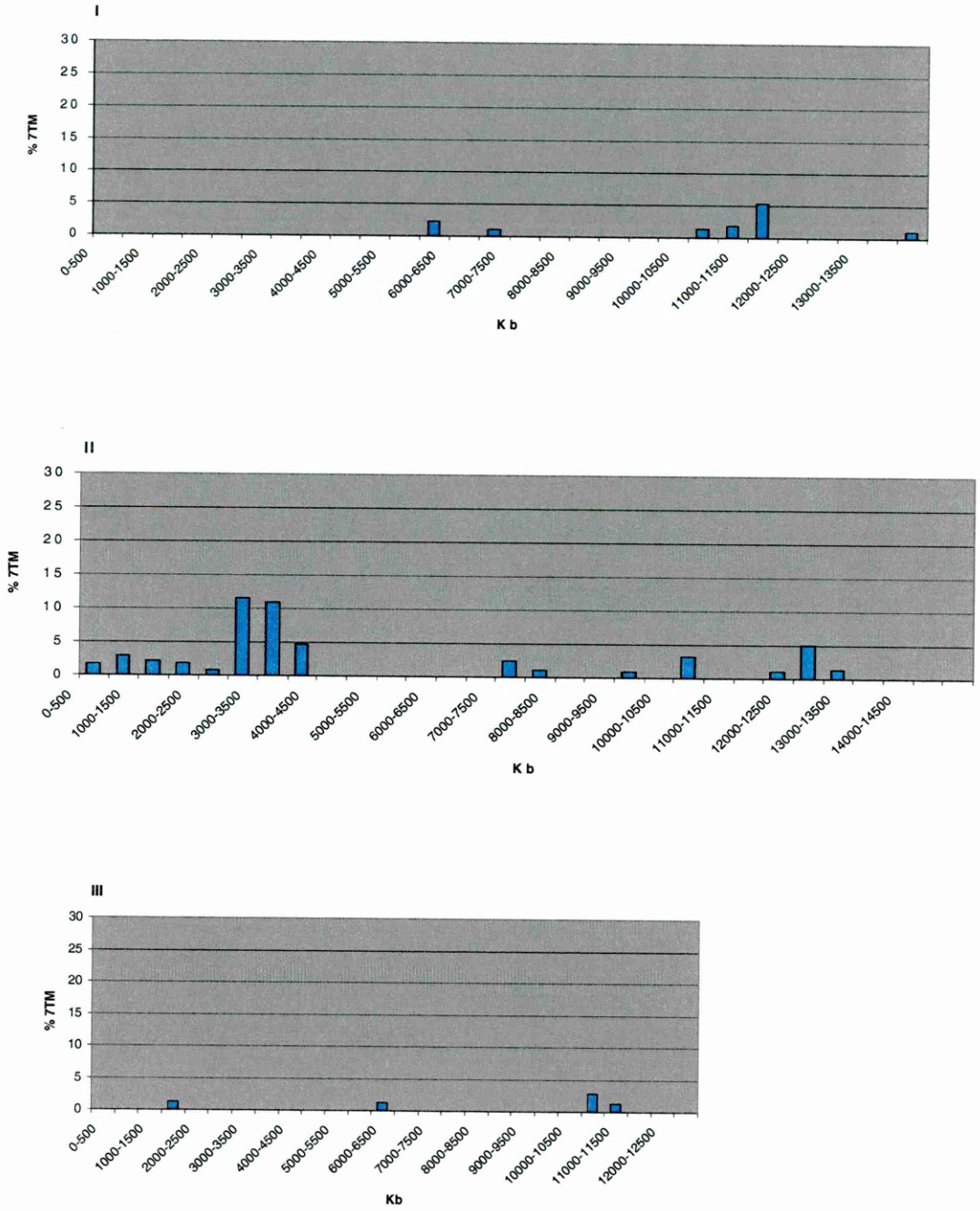
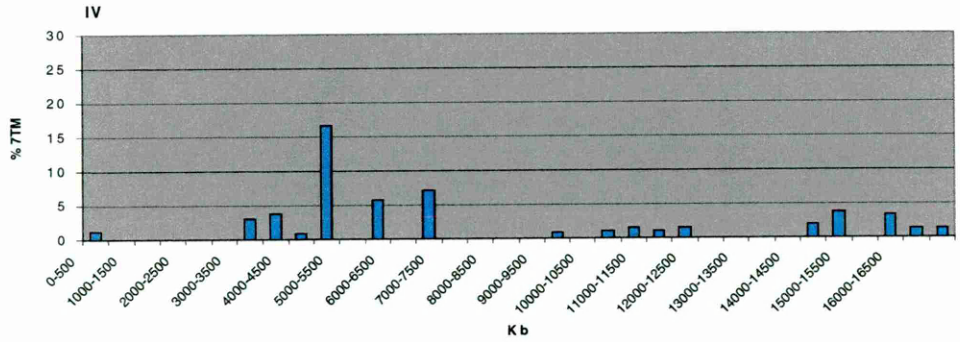
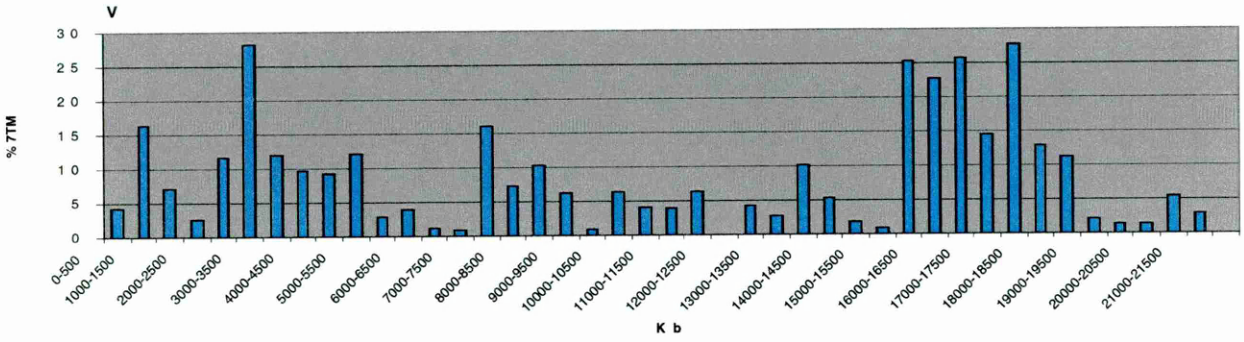



Figure 2-6a: Distribution of putative 7 transmembrane receptors across chromosomes I,II and III. The Y axis indicates the percentage of genes similar to 7 transmembrane receptors in each 500Kb segment across the chromosomes.



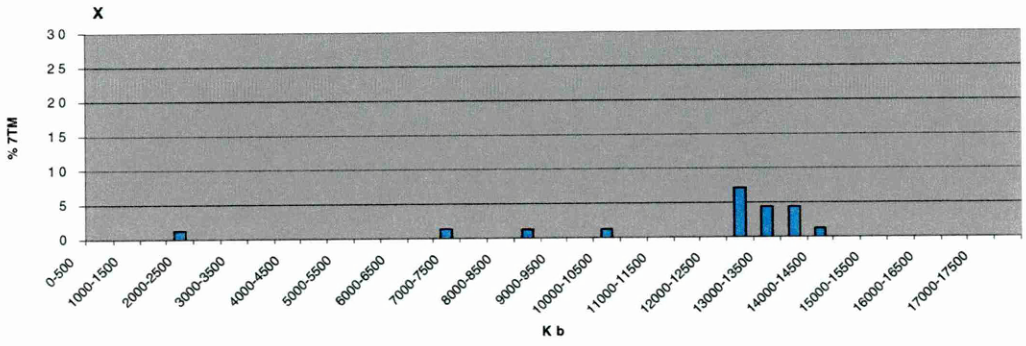




Figure 2-6b: Distribution of putative 7 transmembrane receptors across chromosomes IV,V and X. The Y axis indicates the percentage of genes similar to 7 transmembrane receptors in each 500Kb segment across the chromosomes.

**Table 2-4: Comparison of essential gene content in three genetically balanced regions.**

| Region[1] | | Lethal loci | Predicted total lethal loci | Genes | Size (Mb) | Putative Yeast homologues | Genetic distance (cM) | DNA per cM (kb) | Genes within Gene clusters | Reference |
|---|---|---|---|---|---|---|---|---|---|---|
| I *sDp2* (*egl-30* to *unc-13*) | Arm | 75 | 128 (30.7%) | 417 | 2.052 | 22 (5.23%) | 11.5 | 178.4 | 25 (6.0%) | Johnsen *et al.* submitted |
| | Cluster | 146 | 248 (34.5%) | 718 | 3.339 | 79 (11.0%) | 3.9 | 849.6 | 68 (9.4%) | |
| III *SDp3* (*mec-12* to *mig-10*) | Arm | 7 | 16 (10.2%) | 156 | 0.689 | 3 (1.9%) | 5.5 | 125.3 | 2 (5.1%) | Stewart *et al.* 1998. |
| | Cluster | 93-105 | 207-243 (21.6-25.4%) | 956 | 4.489 | 50 (5.2%) | 4.09 | 1097.6 | 68 (6.4%) | |
| V *ET1* (leftmost end to *mom-2*) | Arm | 52-54 | 70-73 (4.7-4.9%) | 1491 | 6.522 | 62 (4.15%) | 22.6 | 288.6 | 549 (36.8%) | Johnsen and Baillie 1991. |
| | Cluster | 30-36 | 41-49 (8.8-10.5%) | 466 | 1.799 | 21 (4.5%) | 1.4 | 1285.0 | 133 (28.5%) | |

[1] Arm/cluster boundary defined as *unc-73* for I(left), *dpy-27* for III(left) and *StP192* for V(left) as described in Barnes *et al* (1995).

*Discussion*

*C. elegans* chromosomes show both inter-chromosomal and intra-chromosomal variation in their informational content. As had been predicted, to some extent, from previous analyses the central regions of the chromosomes show a greater protein coding density than the flanking arms. Although, the difference in coding density between regions can be small e.g. the left arm of chromosome II is predicted to be 29% coding whilst the central cluster region is 29.68% coding. Overall the range of informational content between chromosomes and their genetically defined compartments is large. For instance, the 6.19MB right arm of chromosome V has only 16.5% of its DNA content predicted as protein coding compared to 32.21% for the 4.98MB central cluster region of chromosome III.

The current estimate of the number of proteins that are coded for by the genome is 19,141. This estimate will undoubtedly change. Initially, this will be due to the completion of the sequence with closure of all gaps. However, the actual number will continue to change in the light of new computational prediction methods and new experimental data. It is likely that the true number will be a source of debate and conjecture for many years.

The gene density on the autosomal arms is higher than had been previously anticipated using calculations based on genetic dissection and EST density [Barnes *et al.* 1995]. Unexpectedly, the left arm of chromosome II and left arm of chromosome V both show gene densities higher than their respective

cluster regions. The remaining autosomal arms all show gene densities less than their adjoining cluster regions. However, the overall trend for *C. elegans* chromosomes remains that recombination rates are higher within the less gene dense autosomal arm regions. Therefore, *C. elegans* differs from other organisms studied such as maize [Civardi *et al.* 1994], human [Ikemura and Wada 1991; Mouchiroud *et al.* 1993] and *Drosophila melanogaster* [Ashburner 1989] where rate of recombinational exchange is positively correlated with higher gene density.

Other intriguing differences between the chromosomes exist. The comparatively high number of tRNA genes on the X chromosome (initially observed by L. Hillier) has yet to be explained. Any reasons why the X chromosome would be a more preferential environment for these genes, either for their creation or their persistence, are not immediately apparent. In the autosomal compartments, tRNA gene density is also remarkably inconsistent, ranging from 2.1 per megabase on the left arm of chromosome I to 19.5 per megabase on the right arm of chromosome V. The tRNAscan-COVE program used to predict the presence of tRNA genes [Lowe and Eddy 1997] also uncovers another phenomenon. In being able to predict low scoring tRNA gene predictions likely to be pseudogenes, a great deal of variance in the percentage of predicted tRNA pseudogenes within each compartment is observed. This ranges from over 61% (94/152) on the right arm of chromosome V to the central clusters of chromosomes III and V which contain no predicted tRNA pseudogenes whilst containing 42 and 37 tRNA genes respectively.

The large number of tRNA genes on the right arm of V is consistent with this chromosome having undergone recent expansion mediated by unequal recombinational events. A recent expansion of this chromosome is suggested by the higher abundance of protein coding gene clusters [this thesis].

We can propose that a genomic region undergoing a heightened rate of recombination mediated gene duplication will accrue a greater proportion of the tRNA genes. In such a scenario, when a gene duplication gives rise to a new tRNA gene then an overabundance of that particular tRNA gene type occurs (assuming selection pressure acting upon the tRNA gene complement remains unaltered). This redundancy now means that any synonymous tRNA gene anywhere in the genome acquiring a deleterious mutation would represent a selectively neutral change (or a beneficial change since this would eliminate the over supply in that particular tRNA species) i.e. any synonymous tRNA gene becomes a potential pseudogene target. This process would ultimately result in tRNA genes accumulating within regions with high rates of gene duplication whilst diminishing in regions with low rates of duplication. This theory may explain the high density of tRNA genes on the right arm of chromosome V. However, such a relationship between tRNA density and gene duplication (as observed through the presence of protein coding gene clusters) elsewhere in the genome is unclear. Obviously, the ability of any genomic region to sequester a larger proportion of the tRNA gene complement through expansion will depend on its initial tRNA gene content in terms of both tRNA type and number.

The lower number of putative yeast homologues and the higher number of putative human homologues on the X chromosome may be indicative of a more ancient period of gene duplication and expansion early in metazoan evolution. This may have been when the X chromosome acquired a large proportion of the tRNA gene compliment.

Regional differences in the EST representation of genes are also apparent within the *C. elegans* genome. Comparison of regions with high EST coverage and regions with a high density of gene clusters [tables 4-1 and 4-3] suggests that regions with a lower density of gene clusters are more highly represented within the EST datasets. One possible explanation of this observation is that genes within clusters may possess fully redundant functions or non-redundant yet overlapping functions. For example, one role of gene duplication in metazoans could be to allow each duplicate to derive a different expression profile, however the total requirement for the gene product and therefore the transcription of both duplicates need not be greater than the previous requirement for the single ancestral gene product. This would mean that overall, individual genes within clusters could be expressed at lower rates than genes with fully unique functionality. Although it should be noted that the extent by which these observations can also be influenced by the characteristics of a relatively small number of gene families has yet to be fully quantified. For instance, many of the gene clusters consist of 7 transmembrane receptors which are expressed at very low levels and rarely represented within EST datasets.

Intron size also shows variation across and between the autosomes and the X chromosome whilst exon size appears to be more consistent across the chromosomes. The higher incidence of repetitive sequences, especially simple tandem and inverted elements, on the autosomal arms may be indicative of the actual physical mechanism by which an intronic expansion has occurred. However, the actual selective pressure(s) causing this phenomenon remains unclear.

One possible advantage of increased intron size is that genes with larger introns will be greater targets for recombination and undergo a higher rate of intra-genic recombination than genes within the cluster regions. The reduced gene density will also result in higher rates of intergenic recombination i.e. a recombination event is more likely to occur between two close gene neighbours on the autosomal arms than within the cluster region.

This effect is further compounded since the autosomal arms also sequester approximately 90% of the recombinational events [Barnes *et al.* 1995]. It should be noted that there is no evidence, as yet, that the larger introns on the autosomal arms have been of subject of expansion. It cannot be ruled out that introns in the central autosomal clusters have also undergone a contraction. This would act to diminish the effect of recombination within the genetic cluster regions.

It should be noted that expansion of the intronic sequence may be a consequence of the heightened recombinational rate within the autosomal arms. The effect could be due to an increased rate of "slippage" and other

recombinational errors. However, the X chromosome displays a relatively linear rate of recombination across its length and also shows an increase in intron size at its extremities. This would suggest that the increase in intron size is not related to an elevated rate of recombination. Although, it should be noted that the X chromosome also displays an increase in its median intron size at its centre, a feature not observed in any of the autosomes. This raises the intriguing possibility that the large X chromosome is the result of an end to end fusion of two chromosomes. Such a fusion may also account for the more linear recombinational pattern observed across the length of the X chromosome. This could be either because the more complex recombinational pattern in the fusion product would tend to more closely resemble a linear pattern or because the normal recombinational patterns for the chromosomal regions are unable to become established. Such a chromosomal fusion may have been a pivotal event in the speciation process of *C. elegans*. However, no other data supports the recent chromosomal fusion outlined above.

The recombinational partitioning of the *C. elegans* genome can be proposed to result in differential rates in gene evolution within the two regions. The autosomal arms undergo higher rates of recombination to produce new genes through duplication and increases in recombination serving to allow an increased rate of their allelic assortment. The increase in intra-genic recombination also allows for an increased rate in the shuffling of allelic determinants. Thus the autosomal arms show properties which may allow the increased rate of duplication and evolution of the genes therein. Since these

regions are therefore more likely to derive and develop new gene functionality, it is possible that these regions play a role as genomic "gene nurseries". As *C. elegans* chromosomes are holocentric possessing multiple kinetochores it is unclear to what extent the relationship between genetic information and recombination will be representative of that in higher monocentric organisms. It should be noted though that during meiosis I in *C. elegans* the pole-ward end of the chromatids perform a function similar to the monocentric centromere [Albertson *et al.* 1997].

Differential rates in the evolution of *C. elegans* genes have been previously reported by Mushegian *et al.* [1998]. In this limited study of 36 orthologous proteins between human, *Drosophila* and *C. elegans* approximately two thirds of these genes where found to have evolved more rapidly in *C. elegans* than their *Drosophila* counterparts. Therefore, if differential rates of evolution in *C. elegans* genes exist then these differences may be related to their chromosomal location. It can be seen from figures 4.3a and 4.3b that the degree of conservation between yeast and human proteins differ markedly across the autosomes. All autosomal arms show a decrease in the density of genes with yeast orthologs compared to their central regions. This suggests that the genes in these regions have undergone divergence more recently consistent with higher rates of evolution in these regions. This effect is much more prominent on the left arm of chromosome II and on both arms of chromosome V.
Chromosome III also shows marked differences between the other chromosomes. It has been shown [this thesis] that chromosome III has fewer

gene clusters than observed on the other autosomes. It is also the smallest chromosome, which may be attributable to the lack of gene cluster forming unequal recombination which have expanded the other chromosomes. Chromosome III, especially in its central cluster region, displays a higher percentage of putative yeast and human orthologs than the other chromosomes. Therefore, the fact that chromosome III has undergone to a reduced extent the processes of gene duplication and subsequent divergence means that it now contains less novelty than other autosomes and may represent a more ancient form of a nematode chromosome. Why this chromosome should have been excluded from the full extent of processes ongoing on the other autosomes is not clear.

A model for *C. elegans* genetic organization can be proposed. The central cluster regions of the autosomes contain the "housekeeping" genes associated with eukaryotic cellular life as indicated by the high orthology of proteins from these regions to proteins within a singled celled eukaryote. These genes will be ancient in origin predating the radiation of the major animal phyla estimated to be 540 to 580 million years ago [Knoll 1992]. Two presumptions can be made about these genes. 1) Since these genes are ancient, natural selection will have already been presented with a vast variety of different allelic assortments. 2) The age of these genes and the fact that they continue to retain sequence similarity with yeast suggests that beneficial mutations continue to arise rarely, if at all, in these genes. Therefore, since the ability for new beneficial alleles in these regions to arise is limited and that a large number of allelic combinations have

already had natural selection applied then it can be argued that recombination will have limited role in these regions. A high recombinational rate in these regions would more likely be deleterious. It will be much more likely that recombination in these regions will provide a less viable allelic assortment than uncover superior combinations i.e. these regions would have a high recombinational load [Charlesworth and Charlesworth 1975]. Therefore, the lowering of the recombinational rate in these regions has the effect of preserving proven and successful combinations of alleles. The central clusters would only require a basal level of recombination sufficient to prevent the accumulation of deleterious alleles.

From the ortholog analysis with yeast and human genes the autosomal arms are more likely to contain genes which have arisen more recently than the radiation of the major animal phyla. The gene cluster analysis shows that the arising of new genes continues to be an ongoing process. Recombination is more likely to uncover beneficial allelic combinations in these regions and so an enhanced rate would be selected for. The increased intron size in these regions will make the genes larger targets for intra-genic recombination. The higher rate of intra-genic recombination would also increase their potential to gain new or improved functionality and evolve.

By partitioning its genome in this way, it can be proposed that *C. elegans* can effectively alter the rates of evolution of genes within each partition by modifying recombinational rates. In some ways, we might consider the *C. elegans* genome to have adopted an "evolutionary bias" by separating and

compartmentalizing those genes which are more likely to evolve into successful variants and genes which are unlikely to be improved upon. Of course, such an evolutionary strategy is based on past success and whilst it may not necessarily be the best strategy in the future it has so far been a successful evolutionary strategy for *C. elegans*.

It is not known whether the modification of recombinational rates via a *rec-1* mediated process is the only way in which *C. elegans* could influence the evolutionary rate of its genes. No clear evidence exists that the mutational rate (except for those mutations caused through unequal recombination) varies over the chromosomes. The higher abundance of repetitive elements on the autosomal arms may be indicative of a higher mutational rate derived from non-recombination related sources. One clue may be derived from the tRNA pseudogene analysis where, overall, the pseudogene rate is much higher on the autosomal arms than on the central autosomal cluster regions.

The X chromosome does not possess a recombinationally lowered central cluster region and its recombinational rate is more uniform across its entire length [Barnes *et al* 1995]. The X chromosome exists hemizygously in males. Males arise spontaneously at a frequency of approximately 1 in 500 animals and natural populations are probably composed predominately of hermaphrodites [Hodgkin 1988]. Therefore, it can be argued that overall, recombinational rates determined in hermaphrodites are not diminished greatly by its role as sex determining numerator. It cannot be precluded, however, that the incidence of males in *C. elegans* was considerably higher in the past. However, the X chromosome has a

smaller proportion of orthologs to yeast genes than the other autosomes. On the autosomes, especially the cluster regions, the proportion of genes currently with human orthologs is less than the proportion of genes with yeast orthologs. This trend is reversed on the X chromosome where more regions show a greater concentration of putative human orthologs than yeast orthologs. A similar scenario is seen on the left arms of chromosome I and IV. These data give may give some clues to the evolution of the *C. elegans* chromosomes. The marked reduction of yeast orthologs on the X chromosome compared to human orthologs suggests that many of these genes have arisen during the evolution of metazoans i.e. they arose after a shared common ancestor with yeast but before the divergence of the major animal phyla. A similar inference could be made for the left arms of chromosome I and IV. Another interpretation on why the X chromosome has less putative orthologs to yeast genes may be due to the fact that it undergoes dosage compensation. If the orthologs to yeast genes serve primarily "housekeeping" functions then we can expect them to be extensively expressed in most cell types and as such may be less suited to existing within a dosage compensated environment.

The genetic amenability of *C. elegans* also shows variability across the chromosomes. Through classical genetic screens the percentage of genes defined ranges from 1.14% of the 1049 genes on the left arm of chromosome II to 8.28% of the 1171 genes on chromosome I. Why does *C. elegans* show such variability in mutationally defined loci? In a cavalier sense, for a gene to continue exist it must exhibit a phenotype which will allow its continued selection.

Therefore all genes and all regions might initially be thought to be equally genetically amenable. Several factors can be invoked in dictating the amenability of a region. Firstly, redundancy of a region will play a role. It has been shown that regions with high proportions of gene clusters show a lowered presence of mutationally defined loci. Since many of genes in gene clusters are similar, strategy they may possess semi-redundant functions i.e. when one is knocked out the other gene functions can compensate in such a way so that the resultant phenotype is subtle.

This work has also shown that the central genetic cluster regions of *C. elegans* are more likely to possess proteins of ancient origin as these regions show a higher degree of homology to genes from the single celled eukaryote *Saccharomyces cerevisiae*. It can be argued that such ancient proteins will be involved in fundamental cellular processes such as metabolism, mitosis and meiosis, the disruption of which may be more likely to elicit an obvious phenotype.

Another interpretation of these data is that the genetic cluster regions accumulate genes which cannot tolerate a high forward mutation rate i.e. genes coding for proteins with a high proportion of amino acids critical for adequate function. As such genes would be constrained in their ability to diverge, we would expect such genes to possess higher rates of similarity with their yeast homologues. This would result in Yeast/*C. elegans* homologues being more readily detected within the genetic cluster regions.

Green *et al.* [1993] also suggested that genetic amenability may also be influenced by transcriptional rate due to the fact that highly expressed genes will be more likely to be highly optimized. Such genes will therefore be more likely to have their functionality perturbed through mutagenesis. This study has shown a positive correlation between EST abundance and the abundance of mutationally defined loci.

The functional class of genes in each region will also effect genetic amenability. As discussed earlier, gene clusters contain a high proportion of seven transmembrane receptors which function as chemoreceptors. This suggests that this is one of the most rapidly evolving gene classes in *C. elegans* [this thesis]. This family of genes is likely to have had some effect on the density of genetically determined genes on chromosome V and also on the left arm of chromosome II. The rapid evolution of the 7 transmembrane receptor family may be due to their involvement in predator/prey interactions and the need to accurately perceive their environment. Such genes by forming the genetic basis of instinct and behavior will be difficult to detect in normal laboratory genetic screens unless specific screens are utilized using particular oderants or ligands. With the complete sequence available, reverse genetic techniques will be able to be used to target specific genes and the protein sequence may give an indication of the putative function and potential ligands so that appropriate phenotypic assays can be made.

The repetitive nature of the *C. elegans* genome repeats also shows regional variance. Overall over 10% of the non-coding DNA sequence and over

7.6% of the entire genome consists of either tandem or inverted repeats. The bias of these repeats toward intronic sequence may be implicated in several processes. For example, alternative splicing can be influenced by repetitive sequence in introns and intron length [Bell *et al.* 1998]. Simple tandem and inverted repetitive elements may also help in the efficiency of the splicing process itself.

Also, as discussed above, repetitive elements may play a role in the general expansion in physical size of a chromosomal region. In general, the autosomal arms in *C. elegans* show a higher abundance of tandem and inverted repeat elements and a lowered gene density. Since 80% and 72% of all the tandem and inverted repeat elements respectively are on the autosomal arms this is a strong indication that the proliferation of repeats can be implicated in the expansion of these regions and the lower gene density.

In *C. elegans* the autosomal arms are the targets of a much higher rate of recombinational exchange. It is therefore possible that the increased rate of recombination is itself responsible for the proliferation of repetitive sequences on the autosomal arms. Errors in the recombinational process may form both tandem and inverted repeat elements. The resulting lowered gene density in these regions would compound the effective rate of recombination for the genes in these regions. In addition, as inverted and tandem repeats in *C. elegans* show a higher relative abundance in intronic sequence, this will tend to increase the intragenic rates of recombination. In maize, recombination takes place within genes and few crossover events take place within intergenic DNA [Civardi 1994].

If a similar bias exists in *C. elegans* then the higher density of intronic repetitive elements may be the result of the concentrated recombinagenic activity within intronic sequence.

It is also interesting to note that percentage of intronic and intergenic repeat sequence is almost identical for both inverted and tandem repeat types. This might be evidence that each repeat type is derived from conversion of the other or produced by the same mechanism

The repeat family elements already determined occupy over 4.5% of the *C. elegans* genome. The various families show intriguing differences in their distributions. One interesting aspect in repeat family distribution are those involving differences on the X chromosome. The X chromosome undergoes a number of processes not associated with normal autosome function. Firstly, the X chromosome undergoes dosage compensation in a similar manner to *Drosophila*, i.e. transcription from the single X chromosome in XO males is increased [Hodgkin 1983]. Secondly, the X chromosome undergoes non-disjunction at a much higher rate than autosomes as part of the mechanism in the spontaneous production of male progeny. In addition, the X chromosome does not display the large variation in recombination across its length that is displayed in autosomes. It is therefore not unreasonable to predict that some of these mechanisms will be mediated through the distribution across the chromosomes of specific DNA sequence elements. Therefore the existence of an X chromosome specific repetitive element would not be unexpected. The counter intuitive result that only repetitive elements have been found with a

strong aversion to X and not vice-versa is intriguing. This suggests that mechanism in the recognition of the dosage compensated chromosome may be through abstention i.e. by dosage compensating the chromosomes that do not actively identify themselves as autosomal. It could also be that an X chromosome determining element has not been detected using the current methodologies e.g. if it is small and/or highly degenerate.

Many other repeat family members show a preference for the highly recombinagenic autosomal arms. It has been speculated for many years that these regions would contain recombination-promoting elements which account for their high level of recombination. CeRep3 (which has also been computationally detected and modeled as CeRep23) has been speculated to be a recombination promoting element due to its higher distribution on the autosomal arms and its lower and more uniform abundance on the X chromosome [Felsenstein and Emmons 1988; Cangiano and LaVolpe 1993; Barnes *et al* 1995]. However, these results also suggest other repeat family elements which are also candidates for recombinational enhancers such as CeRep14, CeRep19 and CeRep24.

# Effects of Gene Expression on Genomic Features

*Introduction*

The higher the overall rate at which a gene is expressed the higher the energetic investment the organism makes in that gene function. Green *et al.* [1993] in their study suggested that moderately expressed proteins show on average a greater degree of sequence conservation over long evolutionary periods than do genes that are rarely expressed. They suggested that this effect is due to higher selective pressures to optimise the protein activity and structure and to minimise potential undesirable interactions with other cellular components. Therefore, if this hypothesis is correct, the transcripts for genes with non-nematoda matches should be represented in the EST datasets at a higher than average rate because sequence perturbation of a conserved gene should be more likely to disrupt function and elucidate a mutant phenotype. We would therefore expect mutationally defined loci to be relatively over-represented in the EST database. With the completion of the *C. elegans* genomic sequence and the availability of large EST datasets such predictions become more directly testable.

In this section the effect of transcription on rate on genomic features is investigated. Several aspects of gene structure can be investigated. For instance, the presence of large introns would theoretically be more energetically

costly than smaller introns in highly expressed genes due to the unnecessary transcription of large amounts of intronic sequence. Other mechanisms, such as splicing would have to be carried out more efficiently in highly expressed genes to avoid a large amount of defective transcripts. Exon length may vary to minimize the number of splicing events allowing the cell to reduce its energetic investment in its splicing apparatus. A reduction in splicing could also enable the message to be processed more rapidly allowing more time for multiple protein translations. Stop codon usage may also be influenced so that translational termination is most efficient in highly expressed genes.

Synonymous codon usage has been found to vary considerably among *C. elegans* genes. Genes which have a relatively unbiased codon usage appear to be expressed at low levels, whilst genes with extremely biased codon usage appear to be expressed at high levels favoring a limited number of translationally optimal codons [Stenico *et al.* 1994]. However, in the study of Stenico *et al.* (1994) the rate of gene expression could not be quantified and relied on anecdotal estimates of expression levels for the 248 genes used in the study. Using EST abundance as a quantitative measure of gene expression we are now able to reappraise the relationship between the usage of optimal codons and expression rate of genes within *C. elegans*.

It can be argued that any selective advantage an organism may derive from changing the genomic features in the above examples must be small, in which case their selection will only have been effected if the long-term evolutionary effective population size of *C. elegans* has been large [Li 1987]. In

other words, has the selective pressure to optimize these genomic components been successful in competing in selective terms with other forms of allelic variation such as beneficial changes in the actual protein sequence, founder effects, mutational biases and hitch-hiking effects caused by the favoring of alleles at other nearby loci?

In order to gauge the expression rate of genes the abundance of transcripts in the EST datasets can be used. Admittedly, this approach has some drawbacks. cDNA libraries used in generating the vast majority of available ESTs are usually subjected to normalization methods to improve the rate of gene discovery. In the cDNA libraries used to generate ESTs in *C. elegans* the abundant transcripts previously identified through EST sequencing were used as hybridization probes [Waterston *et al.* 1992, Kohara 1996]. This allowed previously tagged genes to be identified and removed from the cDNA sets subsequently selected for sequencing. Therefore EST datasets will tend to underestimate the abundance of highly expressed genes. However, in the absence of more accurate measures of gene expression e.g. Serial analysis of Gene Expression (SAGE) [Velculescu *et al.* 1995] EST sequences should remain an adequate guide for the relative abundance of mRNAs within a population.

*Methodology*

Data was taken from *C. elegans* ACEDB release WS6. GFF [Durbin et al. 1997-] data files describing each of the six chromosomes where generated using GIFACE [Durbin and Mieg 1991-].

To determine the exon and intron sizes in genes, only the gene features confirmed by EST data were used in this study. This was to prevent the results being biased by introns and exons predicted only by genefinding programs and no other supporting data. Also, initiating and terminating exons were excluded since their size is not readily quantifiable due to the presence of the untranslated regions (UTRs). The EST dataset used contained 72,451 ESTs derived from Waterston *et al.* [1992], Adams *et al.* [1991] and Kohara [1996]. Stop codons were only studied from genes where at least one of the derived ESTs was obtained from the 3' end of the cDNA clone [Kohara 1996].

The frequency of optimal codons (FOP) was calculated using the program CodonW [Peden and Sharp 1997-] using matrices derived from Stenico *et al.* [1994].

*Results and Discussion*

Of the mutationally defined genetic loci correlated to the sequence map, 72.3% (165/227) had corresponding transcripts represented in the EST dataset. In the entire genome 7362 of the predicted genes are represented in the EST

dataset corresponding to 38.5%. This is a significant discrepancy. Taking into account sampling errors using a binomial distribution at 95% probability these numbers become 72.3% +/- 5.7 and 38.5%+/- <0.5 respectively.

These data show that the genetic loci that currently have been correlated to the sequence map are strongly biased towards representation within the EST dataset. This suggests that amenability of a locus through classical genetic methods is related to its expression rate. As discussed above, this effect may be due to mutations in highly expressed genes being more likely to cause a detrimental effect on the protein function. However, at this stage we cannot preclude the possibility that, overall, genes expressed at low levels are less likely to give a noticeable phenotype in the absence of wild type protein function than more highly expressed genes. Also, genes not represented within the ESTs datasets may not actually be expressed under normal laboratory conditions, e.g. they may be involved in responses to environment stress e.g. heat shock or drought conditions. Therefore, it would be expected that such genes would not elicit a phenotype even if their coding elements were destroyed within mutagenic screens. Alternatively, genes without EST sequences may be more likely to be false predictions, although current analyses would not support this as the major cause of this discrepancy [*C. elegans* Sequencing Consortium 1998]. These results may also be due to biases in the protocols used in correlating genetic loci to the sequence map e.g. transformation rescue experiments may be less successful in cases where the target gene is expressed only at low levels.

The intron and exon size distributions with respect to EST representation are shown in tables 3-1 and 3-2. Overall the correlation coefficient between intron size and EST number is –07844, which indicates that intron size is negatively correlated with EST the significance of this observation exceeding 99% confidence levels.

Table 3-1:Intron size variation with EST representation.

| ESTs | Median size (bp) | Average size (bp) | Number of introns |
|---|---|---|---|
| 1-5 | 58 | 206 | 2970 |
| 6-10 | 62 | 213 | 1525 |
| 11-15 | 55 | 192 | 1113 |
| 16-20 | 52 | 177 | 742 |
| 21-25 | 55 | 200 | 400 |
| 26-30 | 51 | 153 | 339 |
| 31-35 | 51 | 135 | 227 |
| 36-40 | 52 | 142 | 115 |
| 41-45 | 51 | 185 | 85 |
| 46-50 | 51 | 156 | 82 |

Both average intron size and median intron size showed a general decrease with increasing EST representation. However, it can be argued that this relationship is too simplistic. It has already been determined that different chromosomal regions show different median intron sizes [figures 2-1a and 2-1b] and we also know that these regions also differ in their EST representation [Table 2-1]. Therefore it could be possible that the differences observed across the chromosomes with respect to median intron size are due to regional differences in expression rate. However, it is unlikely that the regional differences seen in intron size are due purely to differences in the local expression for two reasons.

Firstly, the intron size differences [figures 2-1a and 2-1b] do not correlate with the different rates of EST representation in each of the genetic compartments [table 2-1]. Secondly, table 3-2 shows that for median intron size chromosomal location may be a more powerful determinant than expression rate. In this table, many of the autosomal arms show an increase in median intron size as compared to the central autosomal regions for genes expressed similar levels. It should also be noted that at the moment of writing the analysis and sequence of autosomal arms was incomplete and that the number of confirmed introns in many of the compartments in table 3-2 is small and therefore will not necessarily be significant. One anomaly is seen with chromosome V, whereby the right autosomal arm shows an increase in median size [figure2-1b] but no associated increase in median size is seen in the EST confirmed intron dataset [table 3-2]. This effect may be due to the high number of 7TM receptors in this region which are almost completely absent in the EST datasets.

Table 3-2: Intron size variation with EST representation within genetic compartments[1].

| # of ESTs | | 1-5 | 6-10 | 11-15 | 16-20 | 21-25 | 26-30 | 31-35 | 36-40 |
|---|---|---|---|---|---|---|---|---|---|
| Chromosome | | | | | | | | | |
| I | L | 183 (45) | 262 (22) | | | | | | |
| | C | 58 (316) | 61 (156) | 60 (105) | 55 (81) | 51 (66) | 58 (67) | 51 (35) | 52 (24) |
| | R | 76 (303) | 127 (192) | 171 (82) | 79 (52) | 106 (30) | 66 (27) | | |
| II | L | 56 (56) | 63 (27) | | | | | | |
| | C | 52 (379) | 55 (197) | 51 (227) | 53 (135) | 50 (70) | 51 (40) | 49 (37) | |
| | R | 71 (111) | 53 (43) | 57 (55) | 102 (30) | | | | |
| III | L | 143 (113) | 57 (31) | 66 (40) | 340 (22) | | | | |
| | C | 56 (110) | 52 (57) | 48 (47) | 49 (24) | 106 (21) | | | |
| | R | 81 (323) | 108 (125) | 76 (89) | 67 (64) | 56 (36) | 48 (51) | | |
| IV | L | 161 (101) | 160 (71) | 107 (20) | | | | | |
| | C | 56 (391) | 57 (238) | 52 (148) | 50 (132) | 57 (69) | 50 (44) | 51 (29) | |
| | R | 62 (133) | 61 (94) | 56 (37) | 78 (30) | | | | |
| V | L | 253 (76) | 62 (32) | 335 (23) | | | | | |
| | C | 52 (258) | 52 (107) | 50 (117) | 48 (85) | 54 (39) | 47 (39) | 68 (26) | 108 (25) |
| | R | 52 (256) | 51 (85) | 52 (79) | 83 (35) | | | | |
| X | | 71 (439) | 69 (274) | 59 (162) | 51 (119) | 79 (75) | 52 (44) | | |

[1]The number of introns in each class is shown in parenthesis. Calculations were only performed when the number of introns was greater than 20.

Although median intron size is influenced primarily by chromosomal location, Table 3-2 also indicates that expression rate does also influence intron size. Of the median exon sizes calculated there are 37 occurrences where the size is lower in more highly expressed genes than those genes with only 1-5 associated ESTs (for the same genetically defined compartment). There are 23 occurrences where the median intron size has increased. The fall in intron size with expression rate may be due to a combination of several factors. For instance, this may be due to a shift towards a shorter intron size in *C. elegans* at which introns are more efficiently and rapidly processed. This effect may also be due to selection towards smaller introns which reduce the amount of intronic sequence transcribed.

Table 3-3: Exon size variation with EST representation

| ESTs[1] | Median size (bp)[2] | | Average size (bp) | |
|---|---|---|---|---|
| | Autosomes | X | Autosomes | X |
| 1-5 | 124 (1198) | 119 (202) | 163 | 138 |
| 6-10 | 139 (733) | 133 (137) | 198 | 173 |
| 11-15 | 157 (581) | 143 (85) | 232 | 221 |
| 16-20 | 190 (423) | 164 (71) | 272 | 209 |
| 21-25 | 172 (218) | 149 (49) | 264 | 233 |
| 26-30 | 211 (200) | 206 (32) | 311 | 333 |
| 31-35 | 235 (145) | | 342 | |
| 36-40 | 174 (78) | | 309 | |
| 41-45 | 210 (57) | | 282 | |
| 46-50 | 262 (53) | | 369 | |

[1]Indicats the range of number of ESTs possessed by the gene from which the exon is derived
[2]The number of exons studied is shown in parenthesis. Calculations were only performed where the number of exons was greater than 20.

Conversely, average and median exon sizes showed an increase in size with increasing EST representation. The actual correlation coefficient for EST abundance and exon size for the 3861 exons under study being 0.223, which is significant above 99% confidence limits. Figures 2-2a and 2-2b show that the chromosomes do not display regional variations in exon size except in the case of the X chromosome where exon sizes are generally lower. Therefore, table 3-3 shows the variance in exon size with expression rate for both autosomes and the X chromosome. Both chromosomal types show increases in exon size with expression rate. In this case, the average exon size is less than 200bp for genes with 10 or fewer ESTs, whilst for genes with 26 or more ESTs the average confirmed exon size exceeds 300bp. Presuming that the coding elements of highly expressed genes are not longer than average then the associated increase in exon size would infer that highly expressed genes possess fewer than average introns. Therefore, these data suggest that transcripts from highly expressed genes undergo fewer splicing events than those from genes expressed at lower rates. The reduced requirement for splicing in highly expressed genes may allow more rapid processing of transcripts and a reduction in aberrant transcripts due to errors in the splicing process. In addition to a lowered requirement for cellular splicing machinery.

The difference in median exon size between highly and poorly expressed genes suggests that the presence of introns within a gene is under selective control at least with respect to its rate of transcription.

The effect of EST abundance on the sequence of the stop codon and subsequent bases was studied. Expression rate influenced the base pair immediately following the stop codon, showing a preference for A at higher transcriptional levels as shown in table 3-4. The influence of stop codon utilization with EST abundance is shown in table 3-8.

Table 3-4: Variance of A in position immediately after the stop codon.

| ESTs[1] | %A | Number of stop codons studied |
|---|---|---|
| 1-10 | 36 | 1134 |
| 11-20 | 45 | 338 |
| 21-30 | 43 | 129 |
| 31-40 | 55 | 64 |
| 41-50 | 71 | 28 |

[1]Indicates the range of number of ESTs possessed by the gene from which the stop codon is derived.

Table 3-5:Stop codon preference with EST abundance

| ESTs | TAG | TAA | TGA | Number |
|---|---|---|---|---|
| 1-10 | 18.3% | 50.7% | 30.2% | 1134 |
| 11-20 | 16.8% | 55.3% | 26.9% | 338 |
| 21-30 | 8.5% | 66.6% | 24.0% | 129 |
| 31-40 | 12.5% | 71.8% | 15.6% | 64 |
| 41-50 | 17.8% | 64.2% | 7.8% | 28 |

In *C. elegans,* overall, stop codon preference was found to be 53.9% TAA, 28.3% TGA and 17.0% TAG. This is remarkably similar to the stop codon variation previously determined in *Saccharomyces cerevisiae* (TAA 55%; TGA 27% and TAG 18%) and a similar bias is observed in *Drosophila melanogaster* [Brown *et al.* 1990].

The preference for TAA as a stop codon increases with expression rate. The base pair immediately following the stop codon also shows variation with expression rate showing a bias towards A as expression increases (see table 3-4). This is consistent with the proposal by Brown *et al.* 1990 that a tetra-nucleotide signals the termination of protein synthesis in eukaryotes; the termination signal consisting of a stop codon and the following nucleotide. Brown *et al* also determined that the preferred stop signal for highly expressed genes in *Saccharomyces cerevisiae* and *Drosophila melanogaster* was TAAG and overall TAA(A/G) and TGA(A/G) are preferred stop signals in eukaryotes. In *C. elegans* the stop signal differs. Overall the preferred stop signals are TAA(T/A) and TGA(T/A) (see table 3-6). The usage of TAA(T/A) increases with expression rate whilst the use of TGA(T/A) shows no increased usage.

The strong preference for TAA(T/A) in highly expressed genes indicates that these represent the optimal termination signals in *C. elegans*. The effect of expression rate on stop signal selection may reflect its influence on the efficiency of the translation termination and subsequent release of the completed polypeptide from the ribosome. In addition, weak stop signals may be mis-read by normal tRNAs [Geller and Rich 1980]. Such termination failure will be more costly in highly expressed genes due to the extra amino acids incorporated into the extended carboxyl termini. In addition, the extended carboxyl terminus may have a detrimental effect on the activity of the resultant protein.

Table 3-6:Variation of 4 most common stop signal consensi with EST abundance

| ESTs | Stops | | | | |
|---|---|---|---|---|---|
| 1-10 | 1134 | TAAT(21.0%) | TAAA(18.0%) | TGAT(12.1%) | TGAA(11.2%) |
| 11-20 | 338 | TAAA(23.6%) | TAAT(21.3%) | TGAA(11.5%) | TAGA(10.0%) |
| 21-30 | 129 | TAAT(30.2%) | TAAA(28.6%) | TGAT(10.8%) | TGAA(9.3%) |
| 31-40 | 64 | TAAA(39.0%) | TAAT(23.4%) | TAGA(9.3%) | TAAC(7.8%) |
| 41-50 | 28 | TAAA(46.4%) | TAGA(14.2%) | TGAA(10.7%) | TAAT(10.7%) |
| Overall | | TAAT(21.7%) | TAAA(21.2%) | TGAT(11.1%) | TGAA(10.9%) |

The frequency of optimal codons (FOP) [Ikemura 1985] was calculated for 2508 genes with 1 or more derived EST sequences. The distribution is shown in figure 3-1 and table 3-10.

Table 3-7: Variation of FOP values with EST abundance

| ESTs | Median | Average | Number of genes |
|---|---|---|---|
| 1-10 | 0.348 | 0.369 | 1155 |
| 11-20 | 0.380 | 0.407 | 273 |
| 21-30 | 0.411 | 0.430 | 108 |
| 31-40 | 0.446 | 0.484 | 55 |
| 41-50 | 0.454 | 0.515 | 27 |

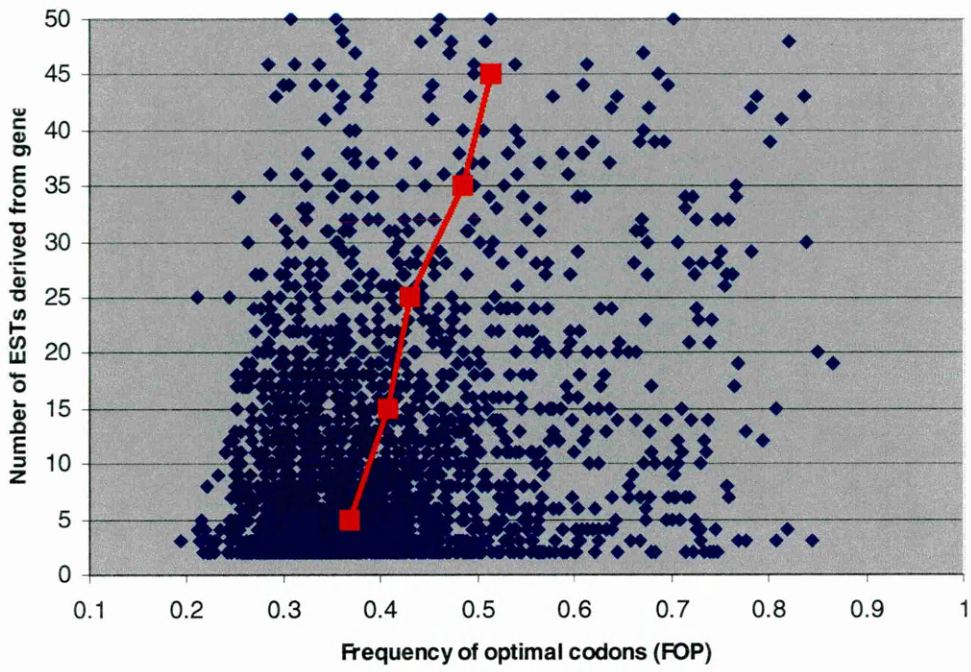



Figure 3-1: Relationship between FOP value and EST abundance. The average FOP values from table 3-11 are also superimposed in orange.

Table 3-10 shows that there is a general association between the FOP value and the level of gene expression determined through EST abundance. Therefore *C. elegans* follows the trend seen in bacteria, unicellular eukaryotes and *Drosophila* where codon usage is associated with transcriptional levels and is consistent with the previous analysis of Stenico *et al.* [1994].

However, gene expression in a differentiated multicellular eukaryote is complex as genes are expressed at different levels in different tissues and at

different times. Other selective pressures on codon usage may exist other than making the translation process more efficient in both speed and accuracy for highly expressed genes [for review see Sharp and Matassi, 1994]. For instance, the selective pressure may be on the speed of translation e.g. for genes involved in regulatory cascades. Also, EST abundance reflects only mRNA abundance in a population. It can be argued that optimal codon usage would also be beneficial in genes which are highly expressed but only within a few cells, as the gains in translational efficiency would make less demands of the individual cells translational machinery. The latter two points would serve to diminish the relationship between EST abundance and optimal codon usage.

The variation of splice site consensus with EST abundance is shown in tables 3-8 and 3-9. Splice donor sites from more highly expressed transcripts show a stronger match to the overall consensus at each position. The splice donor sites from more highly expressed genes also show a significantly stronger match to the overall consensus at positions –8, 2, 3 and 4 as determined through chi-square analysis. Splice acceptor sites show a significantly different base preference (excluding the conserved AG) at all positions studied except –7. At positions –8, 1 and 2 the most preferred base also changes. At position –8 A is weakly preferred over T whilst at the higher expression rate T is preferred. At position 1 A is preferred in the less highly expressed set (41%), whilst in the consensus from more highly expressed transcripts A and G are

Table 3-8: Consensus[1] of splice donor sites from genes with 1-5[2] and >30[3] derived ESTs.

| | | -2 | -1 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| G | 1-5 | 11 | 61 | **100** | 0 | 25 | 10 | 75 | 8 | 13 | 14 |
| | >30 | 12 | **64** | **100** | 0 | 21 | 7 | **77** | 6 | 11 | 13 |
| A | 1-5 | 56 | 18 | 0 | 0 | 58 | 65 | 11 | 20 | 27 | 28 |
| | >30 | **59** | 16 | 0 | 0 | **60** | **68** | 10 | 20 | 26 | 27 |
| T | 1-5 | 18 | 14 | 0 | **100** | 16 | 17 | 12 | 61 | 51 | 48 |
| | >30 | 13 | 13 | 0 | **100** | 18 | 18 | 9 | **63** | **53** | **49** |
| C | 1-5 | 14 | 7 | 0 | 0 | 2 | 8 | 3 | 10 | 10 | 10 |
| | >30 | 16 | 7 | 0 | 0 | 1 | 7 | 3 | 11 | 10 | 11 |
| $\chi^2$ $p$[4] | | 0.0001 | 0.362 | | | 0.003 | 0.012 | 0.048 | 0.132 | 0.35 | 0.67 |

[1] Bold represents the position most similar to the *C. elegans* donor consensus AG/GTAAGTT.
[2] Number of splice sites studied was 4190.
[3] Number of splice sites studied was 683.
[4] Chi-square probability that difference between expression sets is insignificant.

Table 3-9: Consensus[1] of splice acceptor sites from genes with 1-5[2] and >30[3] derived ESTs.

| | | -8 | -7 | -6 | -5 | -4 | -3 | -2 | -1 | 1 | 2 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| G | 1-5 | 7 | 7 | 2 | 1 | 8 | 0 | 0 | **100** | 31 | 15 |
| | >30 | 9 | 6 | 1 | 0 | 6 | 0 | 0 | **100** | 36 | 17 |
| A | 1-5 | 43 | 28 | 6 | 1 | 10 | 3 | **100** | 0 | 41 | 30 |
| | >30 | 38 | 25 | 4 | 0 | 10 | 3 | **100** | 0 | 36 | 28 |
| T | 1-5 | 42 | 58 | 89 | 97 | 67 | 13 | 0 | 0 | 12 | 36 |
| | >30 | 46 | **61** | **92** | **99** | **72** | 13 | 0 | 0 | 11 | 25 |
| C | 1-5 | 8 | 8 | 3 | 2 | 16 | **84** | 0 | 0 | 16 | 19 |
| | >30 | 7 | 8 | 3 | 1 | 11 | **84** | 0 | 0 | 17 | 27 |
| $\chi^2$ $p$[4] | | 0.001 | 0.20 | 0.002 | 0.068 | <0.001 | 1 | | | 0.014 | <0.001 |

[1] Bold represents the position most similar to the *C. elegans* acceptor consensus TTTTTCAG/(AG)(AG).
[2] Number of splice sites studied was 4190.
[3] Number of splice sites studied was 683.
[4] Chi-square probability that difference between expression sets is insignificant. Where necessary values equal to zero were increased to 1 for the chi-square calculation.

equally preferred at 36%. At position +2 T is preferred at low expression rates (36%), at higher rates A is preferred (28%).

*Discussion*

This study presents preliminary evidence using EST data that genes expressed at high levels have proven to be more amenable to classical genetic assay than have genes expressed at a lower rate.

It has also been shown that intron sizes are effected more by their chromosomal location than by the expression rate of their transcripts. Although, within a particular chromosomal compartment intron size is tends to be smaller in more highly expressed genes. In highly expressed genes exon sizes are generally larger. Therefore, gene structure is influenced by expression rate whereby highly expressed genes have larger exons coupled with fewer and smaller introns. The effect on expression appears to be stronger determinant of exon size than on intron size. These data suggest that the efficiency of transcription is more influenced by the number of splicing events required than by the amount of intronic sequence needed to be transcribed.

Expression levels also influence translational termination signals and stop codon preference. In *C. elegans* the preferred termination signal is TAA(A/T), the usage of which increases with the rate of expression. Codon usage is also influenced by expression rate in *C. elegans*, as previously proposed by Stenico *et al.* [1994]. In addition, base preference at splice site consensi is also influenced by expression rate at most positions studied. Therefore, at least for highly expressed genes changes in these genomic features have been able to exert sufficient selective pressure to overcome the processes of random genetic drift.

# Alternative splicing of transcripts in C. elegans

*Introduction*

Alternative splicing has been determined to be involved in an increasing number of genes involved in cellular growth and differentiation and it is common for eukaryotic genes to produce alternatively spliced mRNAs coding for more than one protein product. Many roles for alternate splicing are proposed. Alternate splicing allows a gene to be under the control of multiple promoters allowing the expression of the gene in different cell types, developmental stages and under different environmental conditions. In some cases alternate splicing of transcripts can lead to an inactive protein product and provides another effective mechanism of controlling gene expression. Alternate splicing can also produce proteins where different functionality can be excluded or inserted depending on the exons used.

One of the best studied examples of alternative splicing are the alternative splicing events involved in sex determination in *Drosophila melanogaster* [Baker 1989]. In this process alternative splicing acts as switch producing either active or inactive protein products. In *Drosophila* the *sex-lethal* gene an alternate exon containing a premature stop codon is skipped in females resulting in a functional sex-lethal protein. This in turn regulates the alternative splicing of the

*tra(nsformer)* gene by switching selection of an alternative splice donor in the second exon.

Many examples of alternative transcripts have been shown to exist in *C. elegans*. Examples include *unc-60* [McKim *et al.* 1994] where two actin-depolymerising proteins share an initial coding exon, the only coding element in which is the ATG. In this case, two genes can be regarded as sharing a single promoter. In the case of *unc-33* [Li *et al.* 1992] a second promoter produces a shortened transcript and in *unc-52* [Rogalski *et al.* 1993,1995] a number of different splicing events have been detected including exon skipping and the utilization of alternative carboxyl termini. Five different transcripts have already been identified for *unc-52*. One implication of alternate splicing is that the number of different proteins an organism is able to produce can be significantly higher than the number of genes within the genome.

Mechanisms involved in alternative splicing are not well understood and no *cis*-acting elements or diagnostic sequences have been determined which play a role in this process. Some progress in the genetic dissection of this process has been made. For instance, the production of alternative transcripts for *unc-52* is under the genetic control of the *mec-8* gene. The MEC-8 protein has been predicted to contain RNA binding domains proposed to interact with RNA sequence unique to the nascent *unc-52* transcript. Here the frequency of alternative splicing events in *C. elegans* and repertoire of alternative events used in *C. elegans* is investigated.

*Methodology*

Data was taken from *C. elegans* ACEDB release WS6. GFF [Durbin *et al.* 1997-] data files describing each of the six chromosomes where generated using GIFACE [Durbin and Thierry-Mieg 1991-]. A PERL script utilizing the PERL GFF module [T. Hubbard, unpublished] was used to categorize the alternative events found in the six chromosomes. Alternative splicing events in *C. elegans* are detected due to their notation during the annotation process i.e. if a gene B0393.1 is determined to have an alternative splicing event then each transcript is denoted with a further letter. Therefore the resulting transcripts from this locus will be denoted B0393.1a and B0393.1b. No significance is implied in the alphabetical ordering of transcripts.

Many alternative transcripts have been found and characterized in independent laboratories and the splicing patterns of entire transcripts have been determined. However, the majority of alternative splicing events detected in *C. elegans* genome data have been determined from EST data. Using ESTs as the prime determinant of alternative splicing events has a number of limitations. Detection of a splice variant may be hampered by the fact that alternative transcripts present at low levels will be poorly represented in the EST database (further augmented by cDNA library normalization) and because each EST sequence only describes a small portion of the entire transcript. In addition, since the entire transcript sequence is not derived, the actual number of different processed transcripts cannot be estimated. For example, two different ESTs

derived from different cDNA clones each detect a different alternative splice variant. In this case, the actual number of unique transcripts produced at this locus can vary from between 2 and 4 (i.e. the number of possible unique transcripts varies from between 2 and $2^n$, where n is the number of alternative splices detected by different ESTs). Although this number can be reduced if ESTs can be determined to be derived from the same cDNA clone. Without full length sequencing of transcripts the actual type and number of unique transcripts cannot be determined. Therefore, this study categorizes only the type of alternative splice events detected.

*Results*

For genes that have been studied independently in other laboratories approximately 8% (16/209) were found to produce alternative transcripts. Protocols such as Northern blotting and RACE allowing effective detection of alternative transcripts. This estimate of 8% can also be considered an underestimate because for many genes the possibility of alternative transcripts will not have been investigated. In addition, the transcription profile of their alternative transcripts may make them difficult to determine. In contrast, the total number of genes with alternative splicing events detected by the *C. elegans* sequencing project 190, which represents approximately 1% of the total loci currently predicted in *C. elegans*.

Alternate processing of mRNA transcripts can give rise to a number of different process outcomes. The different classes are outlined in figure 4-1. The characteristics of alternate splicing in *C. elegans* are outlined in table 4-1.

Table 4-1 Summary of alternative splicing in *C. elegans*

| | | % |
|---|---|---|
| Alternatively spliced genes (from total of 19,141) | 190 | 100 |
| Genes with a 5' truncated transcript | 74 | 38.9 |
| Genes with alternative starting exon (B) | 60 | 31.6 |
| Genes with external start site (C) | 6 | 3.2 |
| Genes with internal start site (D) | 19 | 10.0 |
| Genes with a 3' truncated transcript | 58 | 30.5 |
| Genes with alternate terminating exon (F) | 18 | 9.7 |
| Genes with terminating exon extension (G) | 41 | 21.5 |
| Genes with internal splice variations | 61 | 32.1 |
| Genes with alternate exons (I) | 43 | 22.6 |
| Genes with run through exons (L) | 2 | 1.1 |
| Genes with 5' truncated exons (J) | 27 | 14.2 |
| Number of 5' exon truncations | 30 | 15.8 |
| Genes with 3' truncated exons (K) | 24 | 12.6 |
| Number of 3' exon truncations | 24 | 12.6 |

At this point 190 genes have been detected to possess alternatively spliced transcripts. Of these, 74 conceptually produce a protein truncated at the amino terminus, 58 produce a protein truncated at the carboxyl end, and 61 transcripts show an internal variation. Of the internal variations, 30 exon 5' truncations were detected, 22 of which resulted in the deletion of 3 or less codons. In contrast, of the 24 3' exon truncations only 4 resulted in the loss of 3 or less codons. The distribution of the sizes of these truncations is shown in figures 4-2 and 4-3.

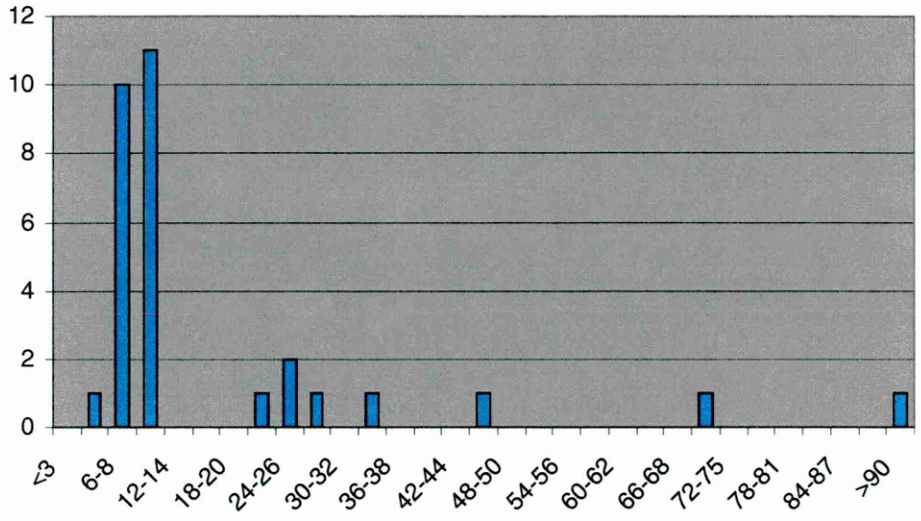



Figure 4-2: Size distribution of 5' exon truncations (base pairs)

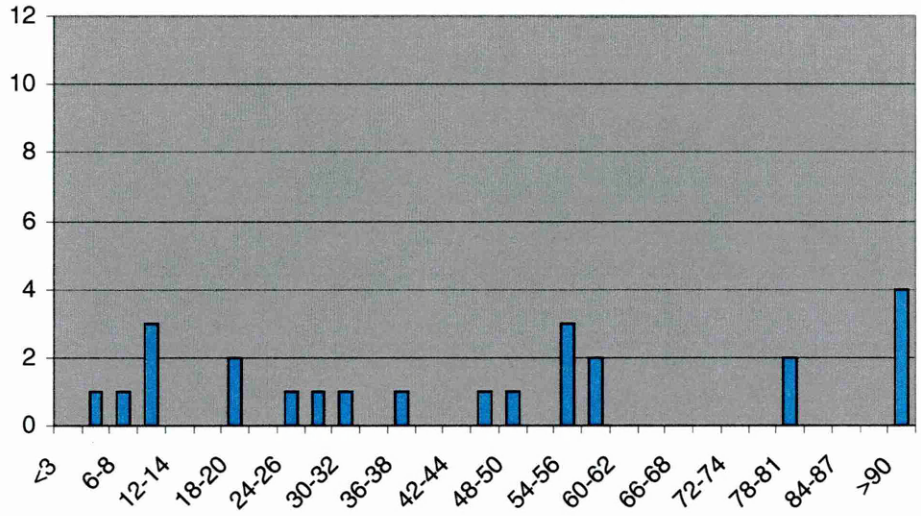



Figure 4-3: Size distribution of 3' exon truncations (base pairs).

*Discussion*

It is clear from the independent study of genetic loci by other laboratories that many more alternative splice variants from the genomic sequence are still to be determined (assuming that the genetic loci so far studied are not biased towards finding those with alternatively spliced transcripts). Many will not be determined without full-length sequencing or restriction analysis of cDNA clones. Other approaches may also be successful in determining the presence of splice variants. For example, a comparative study of the *bli-4* locus in both *C. elegans* and *C. briggsae* indicated the presence of alternative exons which were absent from the current cDNA clones. These putative exons were subsequently confirmed experimentally [Thacker *et al.* 1999].

It is intriguing to note that no single form of alternative splicing predominates in this dataset i.e. the proportion of alternative splicing events giving rise to 3' or 5' truncations or internal exonal variation do not differ vastly. However, we should note that many cDNA clones are not full length, therefore an EST based approach for finding alternate splice variants will be less effective at finding variants at the 5' end of genes than at the 3' end.

Of the number of genes undergoing alternative splicing, the data now suggests that once a gene undergoes an alternative splicing event then the potential to undergo further alternative is increased. Of the 190 alternatively spliced genes 243 different splices are observed [also the ACEDB database as of 5/99 has 233 alternatively spliced genes of which 27genes have 3 or more

transcripts]. If alternatively transcribed genes are only being determined in the genome at a rate of approximately 1%, it could be expected that only 1% of these gene would have a second detected alternative splicing event. Some genes display a number of alternative transcripts, for example *unc-52* [Rogalski *et al.* 1993,1995] and *bli-4* [Thacker *et al.* 1999] have 4 and 9 different transcripts respectively. An explanation for these observations is that alternative splicing involves recognition and binding of a protein (or RNA) complex to the nascent transcript. It could be that once the complex is attracted to and associates with the transcript it becomes much easier for other mutations to arise elsewhere in the transcript which are able to influence splicing. However, the observations are preliminary and may be influenced by other factors and biases. For example, these observations may be biased if second alternative transcripts are determined predominately in genes with large numbers of ESTs.

The different distribution of the truncation lengths (figures 4-2 and 4-3) at the 5' and 3' ends of exons suggest two distinct processes are producing alternatively spliced transcripts. Therefore, it is tempting to speculate that many of the smaller truncations (<=9 base pairs) are merely the result of errors in the splicing process. This would also suggest that the overall mechanisms and signals involved in splice acceptor processing are more error prone than those involved at the splice donor site. 73% (22/30) of the 5' exon truncations lead to the utilisation of a splice site up to three codons away.

The finding that many exon truncations only alter the protein sequence by three or less amino acids suggests some plasticity exists within the splicing

process whereby a nearby but presumably sub-optimal splice site is utilised. In these cases the alternative transcripts will likely confer little or no difference in the functionality of the resultant protein. Although, it should not be forgotten that the insertion or deletion of even a single amino acid at many points in a protein sequence can cause a deleterious effect and that no formal evidence exists that these small changes are necessarily neutral.

Therefore, it is possible that many alternative transcripts will have little immediate biological relevance but this process may have evolutionary significance in providing some variation in protein sequences with minimal cost.

It is also feasible that the actual splicing process is more error prone than would initially appear. In cases where aberrant splicing introduces a frameshift or premature stop codon in the sequence, the message would be eliminated through the nonsense mediated decay system [Pulak and Anderson 1993]. It could be the case that some of the alternative transcripts observed represent the small proportion of erroneously spliced transcripts which fortuitously retain an open reading frame.

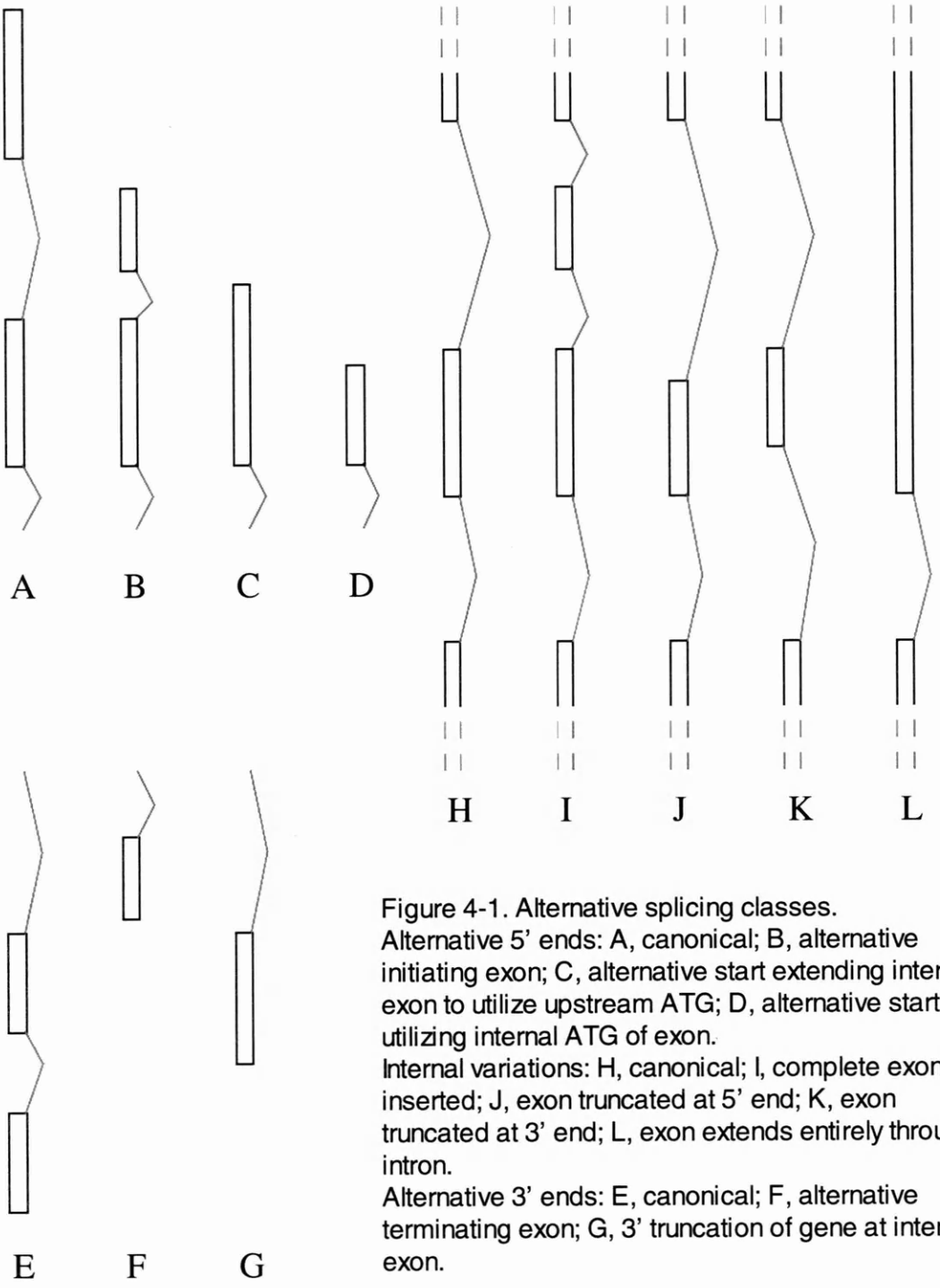



Figure 4-1. Alternative splicing classes.
Alternative 5' ends: A, canonical; B, alternative initiating exon; C, alternative start extending internal exon to utilize upstream ATG; D, alternative start utilizing internal ATG of exon.
Internal variations: H, canonical; I, complete exon inserted; J, exon truncated at 5' end; K, exon truncated at 3' end; L, exon extends entirely through intron.
Alternative 3' ends: E, canonical; F, alternative terminating exon; G, 3' truncation of gene at internal exon.

# Gene Clusters in *C. elegans*

*Introduction*

It is a generally accepted theory that new genes arise through the duplication of pre-existing genes [Ohno 1970]. Gene duplications can arise through a number of means, such as gene conversion, transposition, RNA-mediated exchanges and the duplication of entire chromosomes via polyploidy or aneuploidy. However, it is believed that the principle mode of gene duplication is through unequal crossover events, either between sister or non-sister chromatids [reviewed in Tartof, 1988]. Sturtevant [1925] originally described the phenomenon of unequal crossover. Sturtevant found that the dominant sex linked *Bar* mutation was found to be due to an equal crossover event between non-sister chromatids. However, unequal crossover events due to mis-pairing between sister chromatids is a more common phenomenon. The rate of spontaneous duplication has been determined for two loci in *Drosophila*, *maroon-like* (*ma-1*) and *rosy* (*ry*) [Shapira and Finnerty 1986]. Rates were determined to be $2.7 \times 10^{-6}$ and $1.7 \times 10^{-4}$ per meiosis respectively. Unequal crossover is believed to be a general phenomena in eukaryotes and has been demonstrated in both mice [Harbers *et al.*, 1986] and humans [Jeffreys *et al.* 1988].

Once a gene duplication event has occurred, the presence of two fully redundant copies represents an evolutionary unstable condition. Unless positive

selection pressure exists to maintain both copies then it is probable that one copy will be in time lost through genetic drift. In some cases, the presence of a duplicate has immediate benefits because multiple copies provide enhanced rates of transcription which are beneficial. For example, in the case of ribosomal gene clusters a single copy gene would not be able to be transcribed at the optimal levels required by the cell [Long and Dawid 1980]. If gene duplication occurs through an unequal crossover event then the genes will be in a head-to-tail arrangement, termed a tandem array. Once an unequal crossover event has occurred the likelihood of subsequent mis-pairing and therefore further unequal crossover events is increased causing the tandem array to expand further. The arrangement of genes in a tandem array can be advantageous. Genes in tandem arrays are exposed to the powerful homogenizing mechanism of the combined forces of unequal recombination and gene conversion. This ensures that members of the array do not degenerate and the required number of functional elements is maintained. Such homogenizing effects are responsible for the high degree of conservation between the more than 20,000 tandemly arrayed 5S RNA genes in *Xenopus laevis* [Wolff and Brown 1988].

Genes arranged with tandem arrays can be beneficial, although it has been proposed that in many cases tandem arrays represent an evolutionary unstable formation. In a theoretical treatise, Graham (1995) proposes that the presence of genes in tandem arrays will have two disadvantages. Firstly, the ability of tandem arrays to promote further unequal crossover events will produce chromosomes with increased and reduced array sizes, either of which may

confer detrimental effects. Deletions caused through unequal crossover events are responsible for the sex-linked *bobbed* (*bb*) mutation in *Drosophila* [Frankham *et al.* 1978]. Secondly and more importantly, the homogenizing effects of a tandem array does not engender complexity and the ability of member genes to evolve and acquire a new function is limited whilst within a tandem array. Therefore, it has been argued that in many cases tandem arrays of genes will be temporary genomic features.

Two mechanisms of tandem array breakdown have been proposed. Firstly, the constituent genes can undergo transposition events and disperse throughout the genome. Secondly, the tandem array can form a gene cluster. In this scenario, a gene cluster can be defined as a closely linked collection of related genes, irregularly spaced and often unpredictably mutually inverted. Graham (1997) suggests that the conversion of a tandem array into a gene cluster would proceed primarily through the process of DNA sequence inversion [Hickey et al. 1991] with each inversion event diminishing the effect of unequal recombination on the cluster. Evidence for the conversion between tandem arrays of genes and cluster can be seen in both the U1 and U2 snRNA genes which are tandemly arranged in mice [Dahlburg and Lund 1988; Nojima and Kornberg 1983] but clustered in humans [Bernstein *et al.* 1985; Lindgren *et al.* 1985], although no directionality in the conversion can necessarily be inferred from these observations.

In *C. elegans,* examples of tandem arrays of genes as well as gene clusters have been determined. Ribosomal genes such as the 5S RNA gene are

found in a tandem array of approximately 110 copies [Nelson and Honda 1985]. The six copies of the 16-kD heat shock genes have been determined to be present in 3 pairs [Russnak and Candido 1985; Jones *et al.* 1986], two of which are in an inverted orientation. In this section the abundance of gene clusters in the *C. elegans* genome is investigated.

*Methodology*

Data from the six *C. elegans* chromosomes was derived from the ACEDB data release WS6. Data for each chromosome was obtained in GFF format [Durbin *et al.* 1997-] using GIFACE [Durbin and Mieg 1991-]. Gene clusters were arbitrarily defined as being *n* similar genes within a total of 2*n* genes. Pairs of similar genes were initially identified as being two similar genes within a group of four adjacent genes. These pairs then were used as 'seeds' and extended at their extremities in order to detect larger clusters. Redundant clusters were removed i.e. any gene could only be a member of a single cluster and larger clusters were preferentially retained. Clusters were named after their leftmost constituent on the sequence map. For a cluster to be extended, a *bona-fide* gene cluster was required to be found at each size increment e.g. 8 similar genes arranged in 2 groups of 4 separated by 8 unrelated genes would be counted as 2 clusters, because even though it could satisfy the requirement of 8 similar genes within 16, the intermediate requirement of 6 similar genes in 12 was not met. Similarity was determined using scores derived from the gapped alignment

algorithm BLASTP version 2 [Gish unpublished 1991-1997] and MSPcrunch [Sonnhammer and Durbin 1994] using default parameters.

Gene pairs were compared and percentage similarity values determined using the ALIGN program [Pearson and Lipman 1988].

The genetically defined autosomal compartments were determined using the boundary definitions of Barnes *et al.* [1995]. The clones describing the extent of the central cluster regions were determined as C09D1 to C16C2 for chromosome I; K10G6 to W03C9 for chromosome II; C36E8 to F02A9 for chromosome III; B0547 to ZK897 for chromosome IV and K04A8 to F25H9 for chromosome V.

*Results*

A summary of the clusters found using this methodology using different WUBLASTP similarity thresholds is shown in figure 5-1. The distribution of the clusters across the six chromosomes is shown in figure 5-2.

The distribution of the gene clusters across the chromosomes shows clear non-homogeneity between and across the chromosomes. Chromosome III shows very little evidence of the presence of gene clusters. In total, only 24 clusters of 3 or more similar genes were found (1.8 per megabase). Chromosome V shows the presence of extensive gene clusters across its entire length with 187 clusters of 3 or more similar genes being found (8.7 per megabase). Chromosomes II and I both show an aggregation of gene clusters on a single arm.

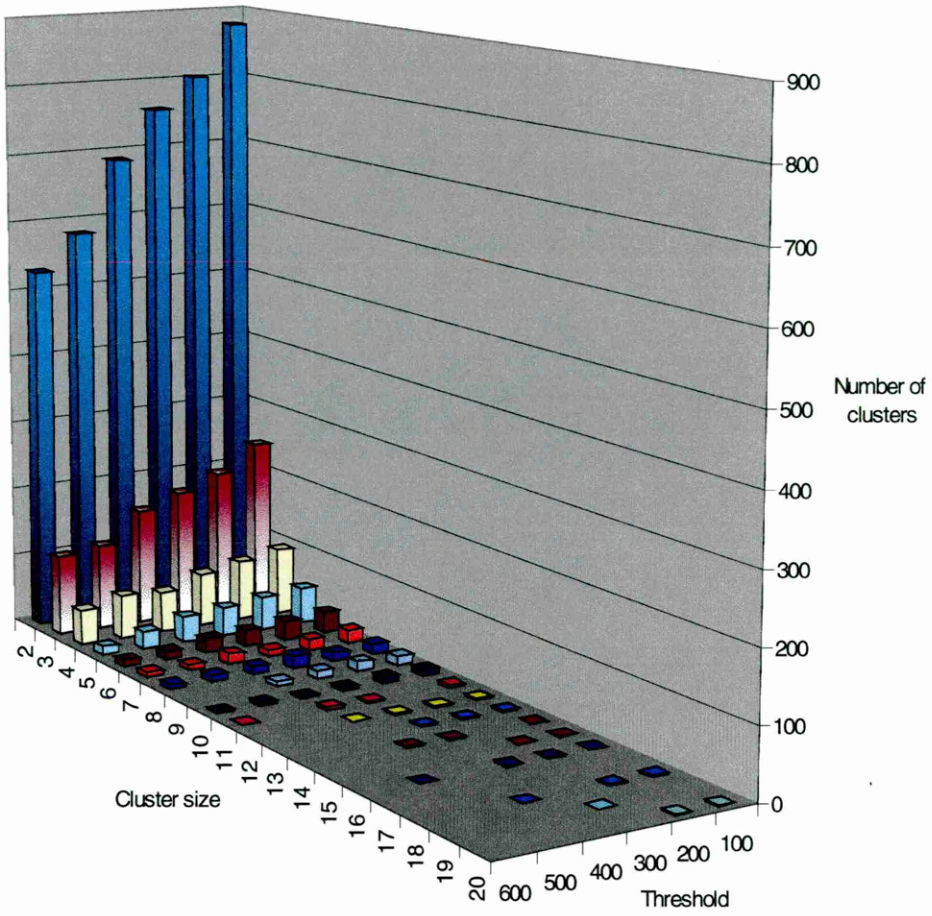



Figure 5-1: The variation of cluster sizes with variation of WUBLASTP threshold

The strandedness of the gene cluster constituents was also was studied (table 5-1). Of the gene clusters of only 2 similar genes, 71% were found to lie on the same strand. For larger clusters (>2), the presence of all the similar genes on the same strand was much higher than expected from random. 204 clusters of 3 similar genes were found, 118 (57%) of which were all on the same

strand (118 +/-14 based on a binomial distribution at 95% confidence). We would expect, if the orientation were random, only 45+/-11 (22%) (i.e.$2*(n/N^2)$, where n is the number of clusters of size N). These data show a significant bias towards cluster expansions forming on the same strand, which is consistent with their expansion being due to unequal cross over events. Gene clusters of nine genes or larger and the orientation of their constituent genes are shown in table 5-1. These larger clusters show grouping of genes with similar orientations within the clusters. 7 of the 33 clusters in table 5-1 have all their constituent genes on a single strand, whilst 12 can be converted to uniform strandedness with a single conceptual inversion event.

The putative functions of the gene cluster constituents from clusters of nine genes and above are also indicated in table 5-1. The largest represented putative function is the 7 transmembrane receptor representing 11 of the 33 clusters. Zinc finger proteins are the second most represented function with 3 clusters of the 33. Nine of the gene clusters cannot as yet be assigned a putative function.

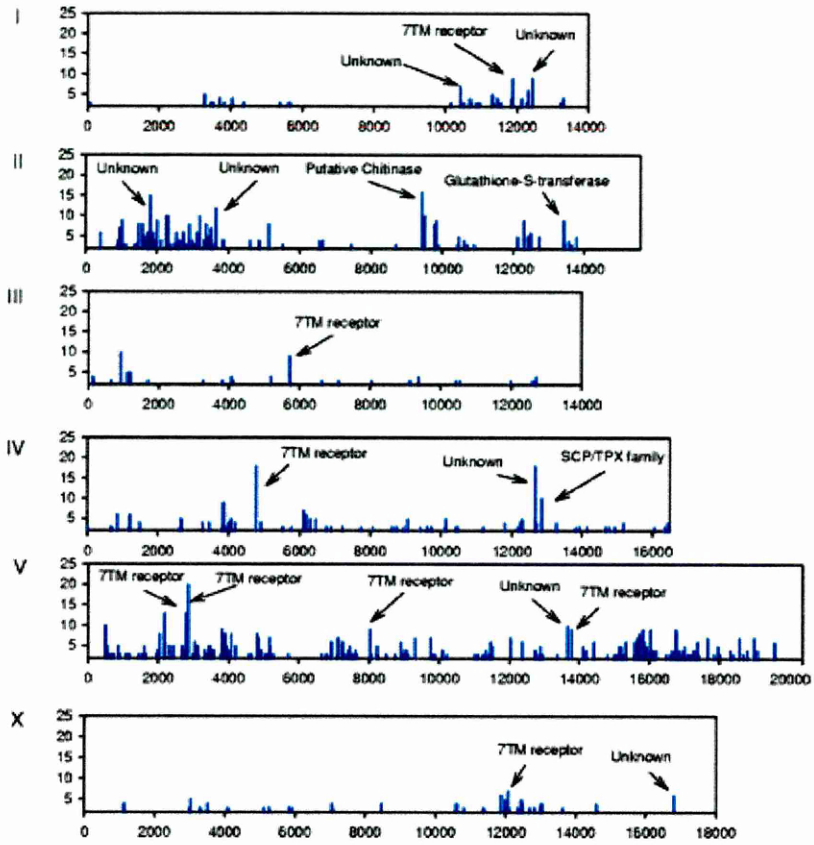



Figure 5-2: Location of gene clusters across the six chromosomes. Only gene clusters with 3 or more members are shown.

Table 5-1: Gene clusters in *C. elegans* with 9 or more constituents[1]

| Size | Cluster Id[2] | Brief Identification | Cluster orientation |
|---|---|---|---|
| 21 | H27D07.5 | Unknown | -----+--++---------- |
| 20 | C32H11.3 | Unknown | +------+-+-++++-+++ |
| 18 | Y17G9A_2.d | 7TM receptor | -+----+--++-++++++ |
| 16 | R09D1.3 | Chitinase | -----+--------++ |
| 15 | T08E11.2 | Unknown | ++++++++------+ |
| 13 | T03D3.6 | 7TM receptor | -+---------- |
| 13 | F44C8.8 | Zn finger protein | --+++++++---- |
| 13 | F43C11.37_a | Unknown | -+++-----++-- |
| 12 | T12B5.8 | Unknown | ---+---+++-+ |
| 12 | F19B10.7 | 7TM receptor | ---+-----+++ |
| 11 | T20B3.8 | C-type lectin binding | +-+-++++--- |
| 11 | K09F6.7 | Zn finger protein | -+--+-+-+++ |
| 11 | K07C6.4 | Cytochrome P450 | +++++----- |
| 11 | F49E11.4 | Testis specific TPX-1 like | ++++++++++ |
| 10 | ZC239.12 | Potassium channel | -+++----- |
| 10 | M02H5.d | Zn finger protein | +++++-++++ |
| 10 | F35E12.7 | Unknown | --------++ |
| 10 | AH6.4 | 7TM receptor | -++++--+-- |
| 9 | ZK938.6 | Chitinase | --------- |
| 9 | ZK1053.4 | Unknown | ++---+--+ |
| 9 | T26E3.9 | 7TM receptor | +++++++++ |
| 9 | T12A2.9 | 7TM receptor | --------- |
| 9 | T06C12.14 | Unknown | +-+-++--+ |
| 9 | F59H6.d | Unknown | ---+----- |
| 9 | F56D6.2 | C-type lectin binding | +++-----+ |
| 9 | F55B12.7 | 7TM receptor | +++++-+++ |
| 9 | F47C10.3 | Nuclear hormone receptor | --------- |
| 9 | F45D11.40_d | Unknown | -+-+-++++ |
| 9 | F37B1.1 | Glutathione S-transferase | --------- |
| 9 | F10A3.8 | 7TM receptor | ------++- |
| 9 | C50E10.6 | 7TM receptor | ----+---- |
| 9 | C09H5.3 | 7TM receptor | --------- |
| 9 | C08E3.5 | 7TM receptor | --------- |

[1] A WUBLASTP/MSPcrunch threshold of 200 was used to determine cluster constituents
[2] Denoted as the leftmost cluster constituent on the chromosome.

For gene pairs their strandedness was studied further. Figures 5-3 and 5-4 show the distribution of the similarity between inverted and tandem gene pairs.

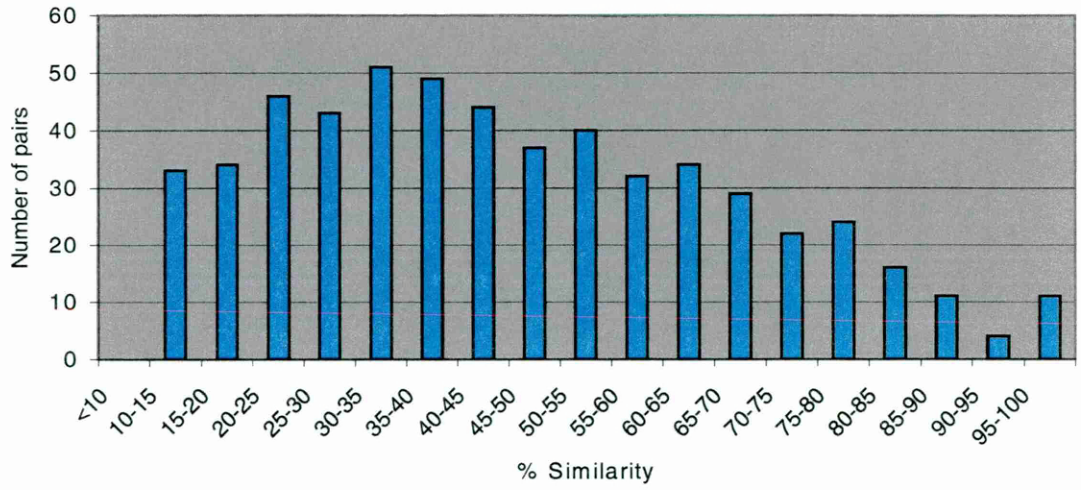



Figure 5-3: Similarity of tandem gene pairs

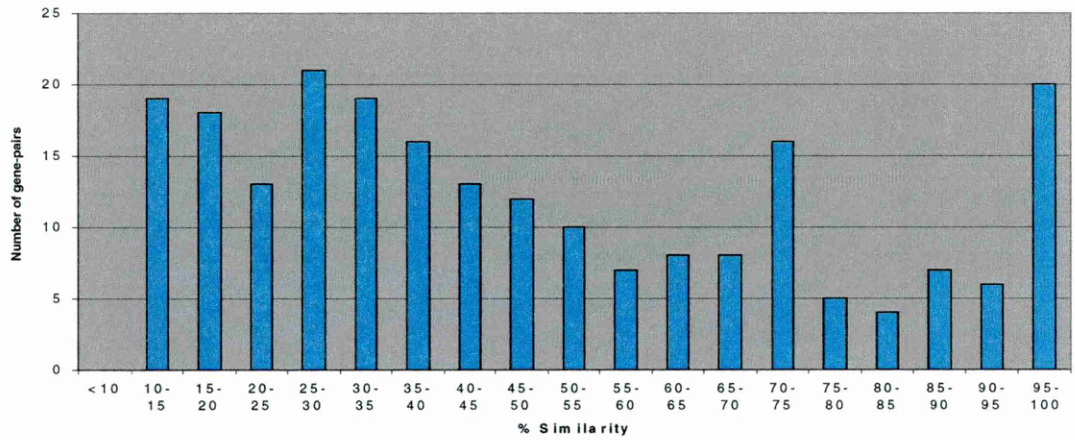



Figure 5-4: Similarity of inverted gene pairs

The occurrence of the two different species of gene pair i.e. tandem (++ or --) and inverted (+- or -+) was also studied with respect to the local recombinational rate (see table 5-2). The results in table 5-2 indicate that inverted gene pairs are more common on the autosomal arms than on the

autosomal central clusters or on the X chromosome. 90% of the recombination events in the *C. elegans* autosomes take place on the autosomal arms. However, although the autosomal arms have statistically significantly more inverted gene pairs than the X chromosome, the difference between the autosomal arms and autosomal cluster region has not quite achieved significance where P=>0.95.

Table 5-2: The proportion of tandem and inverted gene pairs in the genetic compartments[1].

| Type | X chromosome | Autosomal arms | Autosomal cluster region |
|---|---|---|---|
| Tandem | 83% +/- 9.7 (60/72) | 67% +/- 4.1 (314/467) | 74% +/- 5.0 (194/261) |
| | 1 per 287kb | 1 per 99kb | 1 per 122kb |
| Inverted | 17% +/- 8.3 (12/72) | 33% +/- 3.4 (153/467) | 25% +/- 5.4 (67/261) |
| | 1 per 1,430kb | 1 per 303kb | 1 per 475kb |

[1]Error bars derived from binomial distribution and reflect 95% confidence levels.

*Discussion*

The presence of gene clusters in *C. elegans* shows a great deal of variation between the chromosomes. Chromosome V shows extensive clustering of genes across its entire length. This expansion of gene families along its length may be responsible in part for it being the largest chromosome in *C. elegans* being approximately 21.2 megabases in length.

The distribution of gene pairs has been studied in greater depth. The similarity profile of inverted gene pairs (figure 5-4) differs significantly from the

similarity profile of tandem gene pairs (figure 5-3). Inverted gene pairs show a much higher proportion of gene pairs with >95% similarity than tandemly arranged gene pairs.

Initially, when a gene duplication takes place the resulting genes will have a high degree of similarity and redundant functions. This is proposed to represent an evolutionary unstable condition and in the absence for selection pressure for both genes one member will become lost. However, in a subset of gene pairs the members will gain non-redundant (if overlapping) functions and become fixed. The similarity distributions of tandem and inverted gene pairs would initially suggest that inverted repeats have an increased rate in formation due to the higher proportion of highly similar genes. However, a higher rate in the formation of inverted gene pairs is inconsistent with the fact that tandem gene pairs are the most prevalent. A number of explanations can be proposed a) inverted gene pairs are produced by a mechanism independent of that producing tandem gene pairs, the rate of which has increased relatively recently b) inverted gene duplications are produced independently and at a higher rate than tandem gene pairs but are less successful in subsequently gaining non-redundant functions c) inverted gene duplications are more likely to undergo a transposition resulting in the separation of the gene pair d) tandem gene pairs which require high sequence conservation to retain their biological function will be subject to stronger selection pressure to become an inverted pair in order to prevent the promotion of further unequal cross-over events e) inverted gene pairs are more susceptible to gene conversion events. Evidence only exists for the latter

hypothesis. McCormack and Thompson (1990) showed that the frequency of gene conversion is increased between inverted gene pairs. This is thought to be due to the ease of pairing as folding the DNA back upon itself can easily align the two genes.

The presence of gene pairs with respect to local recombination rate has also been studied. Tandem gene pairs are believed to be formed via unequal crossover events. Inverted gene pairs can therefore be formed through the subsequent inversion of one of the constituent genes. The observation that the proportion of inverted gene pairs over tandem pairs appears to be increased in the recombinationally more active autosomal arms is in some ways counter intuitive. However, our results also indicate that inverted pairs are more susceptible to gene conversion events. Therefore, since gene conversion is a phenomenon of the recombination process, inverted gene pairs are more likely to undergo gene conversion in regions where the rate of recombination is high. Therefore, inverted gene pairs are more likely to remain detectable where recombination rate is high. Also, since tandem gene pairs will be more likely to be the subject of further unequal crossover events, gene pairs selection pressure may favor the gene pairs which have undergone an inversion event. Although, it cannot be precluded from this study that inverted gene pairs are formed via an independent mechanism not involving a tandem intermediate and that this process is also sensitive to the local recombinational rate.

If inverted gene pairs are indeed derived from tandem duplications then the rate of sequence inversion within the genome is noteworthy. Overall

approximately 30% of tandem gene pairs undergo an inversion event. This may be a reflection of the general rate of inversion events taking place within the genome as a whole or that specific mechanisms exist to identify tandem duplications and promote sequence inversion. The latter specifically serves to suppress the potential for any subsequent amplification/reduction taking place through unequal crossover.

The behavior of genes observed in gene clusters also gives more clues to the factors affecting genome organisation as a whole. Tandem clusters of genes can be proposed to be deleterious since their presence promotes continued unequal recombination events. This provides selective pressure for tandem arrays of genes to undergo inversion events. However, as inverted gene pairs are more susceptible to undergoing gene conversion events then the ability for their constituent genes to evolve and acquire independent functions is impeded. Therefore, the transposition of a gene from a gene cluster to a region elsewhere in the genome with no nearby close relatives could be considered the best long-term evolutionary strategy. This may help to explain the overall lack of functional organisation within genomes. The apparent random distribution of gene functions within genomes being a selected trait.

The high proportion of 7 transmembrane receptors in the gene clusters indicates that this is the most rapidly evolving gene class within *C. elegans.* The 7 transmembrane receptors which function as G-protein coupled receptors are a fundamental part of the mechanism of oderant perception acting as chemoreceptors to detect oderants and other molecules within the environment.

Through this mechanism of chemosensation, *C. elegans* is able to perceive its environment and modify its behaviour in the presence of various chemical stimuli [for review see Bargmann and Mori 1997]. Since much of the behaviour of *C. elegans* will be in response to chemical signals within the environment, i.e. chemotaxis, we can consider the 7 transmembrane receptors in *C. elegans* as forming the genetic basis of instinct. Already a large number of chemical attractants and repellents have been determined. In addition, the perception pheromones can precipitate entry into the alternative dauer stage as well as stimulating mating behaviour between males and hermaphrodites. The rapid evolution of this protein is therefore not surprising. For *C. elegans* to invade new habitats, the ability to detect new oderants or ligands will be necessary. Also, the ability to detect the presence of predators will require the ability to detect an ever changing repertoire of ligands since we can envisage an "arms race" scenario where predators would constantly be changing to detect perception from their prey. The rapid evolution of the 7 transmembrane receptors may also be more important in *C. elegans* than in higher eukaryotes such as *Drosophila* species and most mammals since *C. elegans* lacks other developed senses such as sight and hearing.

# General Conclusion

The generation of a complete genome for any organism represents a fundamental change in the way its biology can be studied. This thesis has described the initial DNA sequence analysis process as well as determining some properties of the genome in terms of informational content and organisation. Such an analysis can only scratch the surface, the true analysis of the genomic sequence will continue for many years and will involve the whole *C. elegans* community.

Approximately 19,000 protein coding genes have been predicted in the C. elegans. The density of genes differs between chromosomal regions. In addition, the distribution of putative orthologues between *C. elegans* genes and genes from both yeast and human suggests the autosomal arms possess more genes with a more recently evolved functionality. The central regions of the autosomes show a higher density of genes correlated with more ancient functions. This pattern suggests that the autosomal arms behave as "gene nurseries". The higher rates of recombination on the autosomal arms contributing to the higher rates of evolution. Chromosome III is smaller than the other chromosomes and shows a diminished expansion of genes with more recently evolved functions. Chromosome III may therefore be representative of an ancient nematode chromosome. The distribution of tRNA genes within the genome Is also not random. tRNA genes may have become sequestered into

specific parts of the genome, most notably the X chromosome and the right arm of chromosome V, as a consequence of recent expansion events taking place in these regions.

Gene structure also varies between chromosomal regions. The primary variation so far determined is intron size. All autosomes show overall a larger intron size at their extremities than within their central cluster regions. The X chromosome also shows an increase in intron size at its centre. Coupled with the fact that the recombination pattern on the X chromosome differs to that of the autosomes, this raises the possibility that the present X chromosome has arisen from the fusion of two chromosomes.

Correlations have also been made between the ability to detect genes through classical genetic means in different chromosomal compartments. These indicate that genetic loci are more likely to be determined in genes possessing high similarity to genes from distantly related genomes. Genetic loci are also more likely to be determined in highly expressed genes. Regions that have a higher density of gene clusters show a paucity of genetically defined loci. This suggests that the constituent genes will, in many cases, share overlapping or wholly redundant functions.

Gene expression is also able to influence gene structure. Exon sizes are generally larger in highly expressed genes and therefore possess fewer introns. Stop codon preference is influenced by expression rate as well as the subsequent base pair. Highly expressed genes prefer the terminating motif

TAA(T/A). Intron size is less influenced by gene expression rate. The primary determinant affecting intron size being chromosomal location.

Only a small proportion of alternative spliced transcripts have been currently determined. The results also indicate that some splice variants are due to errors in the splicing process, especially at the splice acceptor site.

Gene clusters, i.e. groups of closely related genes, show variation in their distribution across the genome. Chromosome V and the left arm of chromosome II have relatively high densities of gene clusters. Tandem arrays of genes, i.e. where the genes are all in the same orientation, are rare. This indicates that duplicate genes formed through unequal recombination events rapidly undergo subsequent inversion. Such inversion would be beneficial since it prevents the promotion of further unequal recombination events. Alignments between conceptual protein products from pairs of similar genes suggest that inverted gene pairs undergo a higher rate of gene conversion. These results suggest that for closely related genes to diverge and evolve, the best long term evolutionary strategy is transposition to a region of the genome with no closely related genes nearby. In terms of genomic organisation, this means that selective pressure would ensure that gene functions are distributed across the genome in a seemingly random manner.

The availability of sequence data will have profound effects on the approaches and experiments that can now be attempted. An international consortium of laboratories has already been formed to create deletion mutants for each of the predicted genes in *C. elegans.* The current reverse genetics

approach being utilized uses primers designed from the genomic sequence to search a mutant population for an amplified PCR band of less than wild-type size, which is indicative of a deletion. The genomic sequence also allows primers to be designed to amplify a gene sequence which can then be used to eliminate the gene function using RNAi [Fire *et al.* 1998]. Also, being able to view the entire genomic sequence means that the existence of any metabolic or developmental pathway can be queried computationally. Therefore, any potentially redundant pathways can be determined and can both be simultaneously targeted for deletion.

The use of *in-vivo* markers such as green fluorescent protein (GFP) can also heavily exploit the genomic data. Constructs can be made for each gene and the temporal and spatial expression be elucidated. Microarrays are solid supports such as glass onto which DNA preparations such as PCR products or cDNAs are dotted at high density (DeRisi *et al.*1997]. Typically densities in excess of 5,000 individual spots per $cm^2$ are achieved. DNA microarrays can now be constructed containing all *C. elegans* genes. This now allows for the concept of a "reverse northern" whereby a mRNA preparation can be screened for its message content and its relative abundance. This means that the genes undergoing changes in expression, for example, under environmental stress or in response to pharmaceutical assays, can be quickly determined. Using SAGE [Velculescu *et al.* 1995] similar investigations can also be carried out into the differences in composition of mRNA populations.

The genomic sequence also allows the tag sequencing of protein fragments to be quickly correlated with its actual gene. Therefore genes can be correlated to spots on protein gels or to proteins obtained, for example, through immunoprecipitation.

The status of *C. elegans* as the only metazoan with a sequenced genome will hopefully be not long lived. Only with a number of complex genomes sequenced can the true wealth of information be realized and exploited. *Drosophila melanogasta* will be sequenced soon [Rubin 1998] and an accelerated program for the human genome sequence will ensure that its genome will soon also be available [Waterston and Sulston 1998, Wadman 1999]. Other genomes are also in progress such as those of *Arabidopsis thaliana* and rice. Further progress will be made with the sequencing of closely related genomes. The sequencing of *C. briggsae* is underway to provide a comparative study for *C. elegans* whilst attention is turning to the mouse to provide the second sequenced mammalian genome. The mouse being a significant model genetic organism in its own right. The sequencing of a more closely related genome to *C. elegans* will allow us to discern the short term evolutionary events which have taken place since divergence. Comparative analysis will also determine the conserved elements of genes indicative of functional domains in the resultant proteins. However, comparative analysis will likely have a most impact on the ability to detect those sequence elements involved in transcriptional regulation. Now that genome sequencing has made major advances in determining the protein sequence world, the next stage will be

to understand the repertoire of elements involved in the complex role of eukaryotic transcriptional control.

# References

Ahringer, J. (1997). Turn to the worm! *Curr. Opin. Genet. Dev.* 7:410-415.

Albertson D.G., Rose A.M. and Villeneuve A.M. (1997). Chromosome Organisation, Mitosis and Meiosis. pp47-95 in *C. elegans II, Cold Spring Harbor Press, N.Y.*

Alm R.A., Ling L.S., Moir D.T., King B.L., Brown E.D., Doig P.C., Smith D.R., Noonan B., Guild B.C., deJonge B.L., Carmel G., Tummino P.J., Caruso A., Uria-Nickelsen M., Mills D.M., Ives C., Gibson R., Merberg D., Mills S.D., Jiang Q., Taylor D.E., Vovis G.F., Trust T.J. (1999). Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen Helicobacter pylori. *Nature* 397(6715):176-80.

Altschul S.F. Gish, W. Miller E. Myers, E.W. and Lipman D. (1990). Basic local alignment search tool. *J. Mol. Biol.*, 215, 403-410.

Aroian R.V., Levy A.D., Koga M., Ohshima Y., Kramer J.M., Sternberg P.W. (1993). Splicing in *Caenorhabditis elegans* does not require an AG at the 3' splice acceptor site. *Mol. Cell. Biol.* 13(1):626-37

Ashburner M.(1989). Mapping and exchange, pp. 451-501 in *Drosophila*. A laboratory Handbook. *Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.*

Baker, B.S. (1989). Sex in flies: the splice of life. *Nature*. 340:521-524.

Bairoch, A (1993). The PROSITE dictionary of sites and patterns in proteins, its current status. *Nuc. Acids Res.* 21, 3097-3103.

Bairoch, A. and Apweiler, R. (1997). The SWISS-PROT protein sequence data bank and its supplement TrEMBL. *Nuc. Acids Res.* 25(1):31-36.

Bargmann C. I. and Mori I. (1997). Chemotaxis and thermotaxis. pp717-737 in *C. elegans II, Cold Spring Harbor Press, NY.*

Barnes, T.M., Kohara, Y., Coulson, A. and Hekimi, S. (1995). Meiotic recombination, noncoding DNA and genomic organization in *Caenorhabditis elegans*. *Genetics* 141:159-179.

Bateman A., Birney E., Durbin R., Eddy S.R., Finn R.D., Sonnhammer E.L. (1999). Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. *Nucleic Acids Res* 27(1):260-262.

Bell M.V., Cowper A.E., Lefranc M.P., Bell J.I., Screaton G.R. (1998). Influence of intron length on alternative splicing of CD44. *Mol Cell Biol* 10:5930-5941.

Bernardi G. (1993). The isochore organization of the human genome and its evolutionary history--a review. *Gene* 135(1-2):57-66.

Bernstein, L.B., Manser, T., and Wiener, A.M. (1985). Human U1 small nuclear RNA genes: extensive conservation of flanking sequences suggests cycles of gene amplification and transposition. *Molec. Cell. Biol.* 5:2159-2171.

Bird A. and Tweedie S. Transcriptional noise and the evolution of gene number (1995). *Philos Trans R Soc Lond B Biol Sci* 349(1329):249-253.

Birney, E (1997). Dynamite: a flexible code generating language for dynamic programming methods used in sequence comparison. *ISMB* 5:56-64.

Blattner F.R., Plunkett G. 3rd, Bloch C.A., Perna N.T., Burland V., Riley M., Collado-Vides J., Glasner J.D., Rode C.K., Mayhew G.F., Gregor J., Davis N.W., Kirkpatrick H.A., Goeden M.A., Rose D.J., Mau B., Shao Y. (1997). The complete genome sequence of *Escherichia coli* K-12. *Science* 277(5331):1453-1474

Bloom S.E, Delaney M.E. and Muscarella D.E. (1993). Constant and variable features of avian chromosomes. In Manipulation of the avian genome (ed. R.J. Etches and A.M.V. Gibbons), pp. 39-59. *CRC Press, Boca Raton, FL.*

Boguski M.S., Lowe T.M. and Tolstoshev C.M. (1993). dbEST- database for "expressed sequence tags". *Nat. Genet.* 4(4):332-333.

Bonfield, J.K. Smith, K.F. and Staden, R. (1995). A new DNA sequence assembly program. *Nuc. Acids Res.* 23(24):4992-4999.

Borodovsky M., Rudd K.E., Koonin E.V. (1994a). Intrinsic and extrinsic approaches for detecting genes in a bacterial genome. *Nucleic Acids Res.* 22(22):4756-4767.

Borodovsky M., Koonin E.V., Rudd K.E. (1994b*).* New genes in old sequence: a strategy for finding genes in the bacterial genome. *Trends Biochem Sci.* 19(8):309-313.

Brenner S. (1974). The genetics of *Caenorhabditis elegans. Genetics* 77(1):95-104.

Brenner S., Elgar G., Sandford R., Macrae A., Venkatesh B., Aparicio S. (1993). Characterization of the pufferfish (*Fugu*) genome as a compact model vertebrate genome. *Nature* 366(6452):265-268

Bridges C.B. (1935). Salivary chromosome maps with a key to the banding of the chromosomes of *Drosophila melanogaster*. *J. Hered.* 26:60-64.

Brown C.M., Stockwell P.A., Trotman C.N.A. and Tate W.P. (1990). Sequence analysis suggests that tetra-nucleotides signal the termination of protein synthesis in eukaryotes. *Nuc. Acids Res.* 18(21):6339-6345.

Burt D.W., Bumstead N., Bitgood J.J., Ponce De Leon A.F. and Crittenden L.B. (1995). Chicken genome mapping: A new era in avian genetics. *Trends Genet.* 11:190-194.

Cangiano G. and LaVolpe A. (1993). Repetitive DNA sequences located in terminal portion of the *Caenorhabditis elegans* chromosomes. *Nuc. Acids Res.* 21:1133-1139.

*C. elegans* Sequencing Consortium (1998). Genome Sequence of the nematode *C. elegans*: A Platform for Investigating Biology. *Science* 282 (5396): 2012-2018.

*C. elegans* Sequencing Consortium (1991-). The *C. elegans* protein database Wormpep is available from http://www.sanger.ac.uk/C_elegans/Wormpep.

Chalfie M.,Tu Y.,Euskirchen G.,Ward W.W.,Prasher D.C. (1994). Green fluorescent protein as a marker for gene expression. *Science.* 263(5148):802-5.

Charlesworth B. and Charlesworth D. (1975). An experiment on recombinational load in Drosophila melanogaster. *Genetical Research, Cambridge* 25:267-274.

Civardi L, Xia Y, Edwards KJ, Schnable PS and Nikolau BJ. (1994). The relationship between genetic and physical distances in the cloned al-sh2 interval of the *Zea mays* genome. *Proc. Natl. Acad. Sci. USA* 91:8268-8272.

Cole S.T., Brosch R., Parkhill J.,.Garnier T., Churcher C., Harris D., Gordon S.V., Eiglmeier K., Gas S., Barry C.E. 3rd, Tekala F., Badcock K., Basham D., Brown D., Chillingworth T., Connor R., Davies R., Devlin K., Feltwell T., Gentles S., Hamlin N., Holroyd S., Hornsby T., Jagels K., Barrell B.G., *et al (1998).* Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* 393(6685):537-544.

Coulson A.R., Sulston J.E., Brenner S. and Karn J. (1986). Toward a physical map of the genome of the nematode *C. elegans. Proc. Natl. Acad. Sci U.S.A.* 83:7821.

Coulson A., Waterston R., Kiff J., Sulston J., Kohara Y. (1988). Genome linking with yeast artificial chromosomes. *Nature* 335:184-186.

Coulson A., Kozono Y., Lutterbach B., Shownkeen R., Sulston J., Waterston R. (1991). YACs and the *C. elegans* genome. *Bioessays* 13:413-417.

Coulson A., Huynh C., Kozono Y. and Shownkeen R. (1995). The physical map of the *Caenorhabditis elegans* genome. *Methods in Cell Biology,* 48:533-550.

Cross S.H. and Bird A.P. (1995). CpG islands and Genes. *Curr. Opin. Genet. Devel.* 5:309-314.

Dahlberg, J. and Lund, E. (1988). The genes and transcription of the major small nuclear RNAs. Structure and function of the major small nuclear ribonucleoprotein particles edited by Birnstiel, M.L. *Berlin:Springer-Verlag.*

Dear S., Durbin R., Hillier, L., Marth G., Thierry-Mieg J. and Mott R. (1998). Sequence Assembly with CAFTOOLS. *Genome Res.* 8:260-267.

DeRisi J.L., Iyer V.R., Brown P.O. (1997). Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale. *Science.* 278(5338):680-686.

Durbin R. and Thierry-Mieg J. (1991-). A *C. elegans* Database. Documentation , code and data available via anonymous FTP from ftp.sanger.ac.uk.

Durbin R. *et al.* (1997-). GFF documentation available from http://www.sanger.ac.uk/Users/rd/gff.shtml.

Eddy, S.R. (1995-). The HMMER package for using profile hidden Markov models. Code and documentation available from http://hmmer.wustl.edu/.

Eddy, S.R. and Durbin, R. (1994). RNA sequence analysis using covariance models. *Nucl. Acids. Res.* 22:2079-2088.

Elgin S.C.R. (1996). Heterochromatin and gene regulation in *Drosophila.* Curr Opin *Genet Dev.* 6(2):193-202.

Etzold T., Ulyanov A. and Argos P. (1996). SRS: Sequence retrieval system for molecular biology databanks. *Methods in Enzymology* 266:114-128.

Ewebank J.J., Barnes T.M., Lakowski B., Lussier M. and Hekimi S. (1997). Strucural and functional conservation of the *C. elegans* timing gene *clk-1.* *Science* 275:980-983.

Felsenstein K.M. and Emmons S.W. (1987). Strucutre and evolution of a family of interspersed repetitive DNA sequences in *C. elegans. J. Mol. Evol.* 25:230-240.

Fichant G.A. and Burks C. (1991). Identifying potential tRNA genes in genomic DNA sequences. *J. Mol. Biol.* 220:659-671.

Fire A., Xu S.Q., Montgomery M.K., Kostas S.A., Driver S.E., Mello C.C. (1998). Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans. Nature* 391(6669):806-811.

Fleischmann R.D., Adams M.D., White O., Clayton R.A., Kirkness E.F., Kerlavage A.R., Bult C.J., Tomb J.F., Dougherty B.A., Merrick J.M., *et al.* (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269(5223):496-512.

Frankham R., Briscoe D.A. and Nurthen R.K. (1978). Unequal crossing over at the rRNA locus as a source of quantitative genetic variation. *Nature* 272:80-81.

Geller A.I. and Rich A.P. (1980). A UGA termination suppression tRNATrp active in rabbit reticulocytes. *Nature* 283:41-46.

Goffeau A., Barrell B.G., Bussey H., Davis R.W., Dujon B., Feldmann H., Galibert F., Hoheisel J.D., Jacq C., Johnston M., Louis E.J., Mewes H.W., Murakami Y., Philippsen P., Tettelin H., Oliver S.G. (1996). Life with 6000 genes. *Science* 274(5287):546, 563-567.

Graham G.J. (1995). Tandem Genes and Clustered Genes. *J. Theor. Biol.* 175:71-87.

Green P., Lipman D., Hillier L., Waterston R., States D. and Claverie J-M. (1993). Ancient Conserved Regions in New Gene Sequences and the Protein Databases. *Science* 259:1771-1716.

Harbers K., Soriano P., Muller U., Jaenisch R. (1986). High frequency of unequal recombination in pseudoautosomal region shown by proviral insertion in transgenic mouse. *Nature* 324(6098):682-685.

Hickey D.A., Bally-Cuif L., Abukashawa S., Payant V. and Benkel B.F. (1991). Concerted evolution of duplicated protein-coding genes in *Drosophila*. Proc. *Natl. Acad. Sci. U.S.A.* 88:1611-1615.

Hodgkin J (1983). X chromosome dosage and gene expression in Caenorhabditis elegans: Two unusual dumpy genes. *Mol. Gen. Genet.* 192:452-458.

Hodgkin J (1998). Sexual Dimorphism and Sex Determination pp243-280 in The Nematode *Caenorhabditis elegans, Cold Spring Harbor Press, NY.*

Holmquist G.P. (1987). Role of replication time in the control of tissue specific expression. *Am. J. Hum. Genet.* 40:151-173.

Ikemura T. (1985). Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.* 2:13-34.

Ikemura T. and Wada K. (1991). Evident diversity of codon usage patterns of human genes with respect to chromosome banding patterns and chromosome numbers: relation between nucleotide sequence data and cytogenetic data. *Nucleic Acids Res.* 19:4333-4339.

Jeffreys A.J., Royle N.J., Wilson V., Wong Z. (1988). Spontaneous mutation rates to new length alleles at tandem-repetitive hypervariable loci in human DNA. *Nature* 332(6161):278-281.

Johnsen R.C. and Baillie D.L. (1991). Genetic analysis of a major segment [LGV(left)] of the genome of *Caenorhabditis elegans. Genetics* 129(3):735-752.

Jones D., Russnak R.H., Kay R.J., Candido E.P.M. (1986). Structure, expression, and evolution of a heat-shock gene locus in *Caenorhabditis elegans* that is flanked by repetitive elements. *J. Biol. Chem.* 261:12006-12015.

Kaufman B.P. (1939). Induced chromosome rearrangements in *Drosophila melanogaster. J. Hered.* 30:178-190.

Kohara Y. (1996). [Large scale analysis of *C. elegans* cDNA] [article in Japanese]. *PNE Protein Nucleic Acid Enzyme* 41(5):715-720.

Knoll A.H. (1992). The early evolution of eukaryotes: a geological perspective. *Science* 256(5057):622-627.

Koonin E.V., Bork P., Sander C. (1994). Yeast chromosome III: new gene functions. *EMBO J* 1994 Feb 1;13(3):493-503.

Koop B.F., Rowen L., Wang K., Kuo C.L., Seto D., Lenstra J.A., Howard S., Shan W., Deshpande P., Hood L. (1994a). The human T-cell receptor TCRAC/TCRDC (C alpha/C delta) region: organization, sequence, and evolution of 97.6 kb of DNA. *Genomics* 19(3):478-493.

Koop B.F. and Hood L. (1994b). Striking sequence similarity over almost 100 kilobases of human and mouse T-cell receptor DNA. *Nat. Genet.* 7(1):48-53.

Krause M. and Hirsh D. (1987). A trans-spliced leader sequence on actin mRNA in *C. elegans. Cell*, 63:753-761.

Kunst F., Ogasawara N., Moszer I., Albertini A.M., Alloni G., Azevedo V., Bertero M.G., Bessieres P., Bolotin A., Borchert S., Borriss R., Boursier L., Brans A., Braun M., Brignell S.C., Bron S., Brouillet S., Bruschi C.V., Caldwell B., Capuano V., Carter N.M., Choi S.K., Codani J.J., Connerton I.F., Danchin A, *et al.* (1997). The complete genome sequence of the gram-positive bacterium *Bacillus subtilis. Nature* 390(6657):249-256.

Li, W.H. (1987). Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. *J. Mol. Evol.*, 24:337-345.

Li W., Herman R.K. Shaw J.E. (1992). Analysis of the *Caenorhabditis elegans* axonal guidance and outgrowth gene *unc-33. Genetics* 132:675-689.

Lindgren V., Ares Jr. M., Weiner A.M. and Francke U. (1985). Human genes for U2 small nuclear RNA map to a major adenovirus 12 modification site on chromosome 17. *Nature* 314:115-116.

Long E.O. and Dawid I.B. (1980). Repeated genes in eukaryotes. *Ann. Rev. Biochem.* 49:727-764.

Lowe T.M. and Eddy S.E. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nuc. Acids Res.* 25:955-964.

Matzke M.A., Varga F., Berger H., Schernthaner J., Schweizer D., Mayr B. and Matzke A.J. (1990). A 41-42 bp tandemly repeated sequence isolated from nuclear envelopes of chicken erythrocytes is located predominately on microchromosomes. *Chromosoma* 99:131-137.

McCombie W.R., Adams M.D., Kelley J.M., Fitzgerald M.G., Utterback T.R., Khan M., Dubnick M., Kerlavage A.R.,Venter J.C. and Fields C. (1992). *Caenorhabditis elegans* expressed sequence tags identify gene families and potential disease gene homologues. *Nat Genet.* 1(2):124-131.

McCormack W.T. and Thompson C.B. (1990). Chicken Ig(L) variable region gene conversions display pseudogene donor preference and 5' to 3' polarity. *Gene* 94:263-272.

McKim K.S., Matheson C.. Marra M.A., Wakarchuk M.F., Baillie D.L. (1994). The *Caenorhabditis elegans unc-60* gene encodes proteins homologous to a family of actin binding proteins. *Mol. Gen. Genet.* 242:346-357.

McQueen H.A., Giorgia S. and Bird A.P. (1998). Chicken microchromosomes are hyperacetylated, early replicating and gene rich. *Genome Research* 8:621-630.

Metzstein M.M., Hengartner M.O., Tsung, N., Ellis, R.E. and Horvitz H.R. (1996). Transcriptional regulator of programmed cell-death encoded by *Caenorhabditis elegans ces-2. Nature* 382:545-547.

Mouchiroud D., D'Onofrio G., Aissani B., Macaya G., Goutier C. *et al.* (1992). The distribution of genes in the human genome. *Gene* 100:181-187.

Mott R. (1997). EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. *CABIOS* 13(4):477-478.

Mushegian A.R., Garey J.R., Martin J. and Liu L.X. (1998). Large-Scale Taxonomic Profiling of Eukaryotic Model Organisms: A Comparison of Orthologous Proteins Encoded by the Human, Fly, Nematode and Yeast Genomes. *Genome Research* 8:590-598.

Naclerio G., Cangiano G., Coulson A., Levitt A., Ruvolo V. and LaVolpe A. (1992). Molecular and genomic organisation of clusters of repetitive DNA sequences in *Caenorhabditis elegans*. *J. Mol. Biol.* 226:159-168.

Needleman S. and Wunsch C. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48(3):443-453.

Nelson D.W. and Honda B. (1985). Genes encoding for 5S ribosomal RNA of the nematode *Caenorhabditis elegans*. *Gene* 38:245-251.

Nielsen C. (1995). Animal evolution. Interrelationships of the living phyla. *Oxford University press, Oxford, UK.*

Nojima H. and Kornberg R.D. (1983). Genes and pseudogenes for mouse U1 and U2 small nuclear RNAs. *J. Biol. Chem.* 258:8151-8155.

Ohno S. (1970). Evolution by gene duplication. *New York:Springer-Verlag.*

Oliver S.G., van der Aart Q.J., Agostoni-Carbone M.L., Aigle M., Alberghina L., Alexandraki D., Antoine G., Anwar R., Ballesta J.P., Benit P., *et al* (1992). The complete DNA sequence of yeast chromosome III. *Nature* 357(6373):38-46.

Pearson W.R. and Lipman D.J. (1988). Improved Tools for Biological Sequence Comparison. *Proc. Natl. Acad. Sci.* USA 85:2444-2448.

Peden J and Sharp P. (1997-). CodonW: code and source available via anonymous ftp from molbiol.ac.uk.

Pulak R. and Anderson P. (1993). mRNA surveillance by the *Caenorhabditis elegans smg* genes. *Genes Dev.*, 7(10):1885-1897.

Robertson H.M. (1998). Two large families of chemoreceptor genes in the nematodes *Caenorhabditis elegans* and *Caenorhabditis* briggsae reveal extensive gene duplication, diversification, movement, and intron loss. *Genome Res* 8(5):449-463.

Rogalski T.M., Willaims B.D., Mullen G.P. and Moerman D.G. (1993). Products of the *unc-52* gene in *Caenorhabditis elegans* are homologous to the core protein of the mammalian basement membrane heparan sulfate proteoglycan. *Genes Dev.*, 7(8):1471-1484.

Rogalski T.M., Gilchrist E.J., Mullen G.P. and Moerman D.G. (1995). Mutations in the *unc-52* gene responsible for body wall muscle defects in adult *Caenorhabditis elegans* are located in alternatively spliced exons. *Genetics* 139(1):159-169.

Rubin G.M. (1998). The *Drosophila* genome project: a progress report. *Trends Genet.* 14(9):340-3

Russnak R.H. and Candido E.P.M. (1985). Locus encoding a family of small heat-shock genes in *Caenorhabditis elegans*: Two genes duplicated to form a 3.8-kilobase inverted repeat. *Mol. Cell. Biol.* 5:1268-1278.

*Saccharomyces cerevisiae* genome sequencing consortium (1997). The nucleotide sequence of Saccharomyces cerevisiae chromosome XVI. *Nature*, 387(6632 suppl):103-105.

Shapura S.K and Finnerty V.G. (1986). The use of genetic complementation in the study of eukaryotic macromolecular evolution: rate of spontaneous gene duplication at two loci of *Drosophila melanogaster*. *J Mol Evol* 23(2):159-167.

Sharp P.M. and Matassi G. (1994). Codon usage and genome evolution. *Curr. Op. Genet.* 4:851-860.

Sidow A. and Thomas W.K. (1994). A molecular evolutionary framework for eukaryotic model organisms. *Curr. Biol.* 4:593-603.

Smit A.F. (1996). The origin of interspersed repeats in the human genome. *Curr Opin Genet Dev* 6(6):743-748.

Sonnhammer E.L.L. (1996). Classification of protein domain families for genomic sequence analysis. *Ph.D. thesis, The Open University, UK.*

Sonnhammer E.L.L. and Durbin R. (1994). A workbench for large-scale sequence homology analysis. *Comput Appl Biosci* 10(3):301-307.

Sonnhammer E.L., Eddy S.R. and Durbin R. (1997). Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins* 28(3):405-420.

Spieth J., Brooke G., Kuersten S., Lea K. and Blumenthal, T. (1993). Operons in *C. elegans*: polycistronic mRNA precursors are processed by trans-splicing of SL2 to downstream coding regions. *Cell*, 73:521-532.

Stenico A., Lloyd A.T. and Sharp P.M. (1994). Codon Usage in *Caenorhabditis elegans*: delineation of translational selection and mutational biases. *Nuc. Acids Res.* 22(13):2437-2446.

Stewart H.I., O'Neil N.J., Janke D.L., Franz N.W., Chamberlin H.M., Howell A.M., Gilchrist E.J., Ha T.T., Kuervers L.M., Vatcher G.P., Danielson J.L., Baillie D.L.

(1998). Lethal mutations defining 112 complementation groups in a 4.5 Mb sequenced region of *Caenorhabditis elegans* chromosome III. *Mol Gen Genet* 260(2-3):280-288.

Stoesser G., Moseley M.A., Sleep J., McGowran M., Garcia-Pastor M., Sterk P. (1998). The EMBL nucleotide sequence database. *Nucleic Acids Res* 26(1):8-15.

Sturtevant A.H. (1925). The effects of unequal crossing over at the Bar locus in *Drosophila*. Genetics 10:117-147.

Sulston J., Du Z., Thomas K., Wilson R., Hillier L., Staden R., Halloran N., Green P., Thierry-Mieg J., Qiu L., *et al* (1992). The *C. elegans* genome sequencing project: a beginning. *Nature* 356:37-41.

Tartof K.D. (1988). Unequal crossing over then and now. *Genetics* 120(1):1-6.

Thacker C., Marra M.A., Jones A., Baillie D.L., Rose A.M. (1998). Functional genomics in Caenorhabditis elegans: An approach involving comparisons of sequences from related nematodes. *Genome Res* 9(4):348-359.

Tomb J.F., White O., Kerlavage A.R., Clayton R.A., Sutton G.G., Fleischmann R.D., Ketchum K.A., Klenk H.P., Gill S., Dougherty B.A., Nelson K., Quackenbush J., Zhou L., Kirkness E.F., Peterson S., Loftus B., Richardson D.,

Dodson R., Khalak H.G., Glodek A., McKenney K., Fitzegerald L.M., Lee N., Adams M.D., Venter J.C., *et al.* (1997). The complete genome sequence of the gastric pathogen Helicobacter pylori. *Nature* 388(6642):539-47

Troemel E.R., Chou J.H., Dwyer N.D., Colbert H.A., Bargmann C.I. (1995). Divergent seven transmembrane receptors are candidate chemosensory receptors in *C. elegans. Cell* 83:207-218.

Turner B.M. (1993). Decoding the nucleosome. *Cell* 75:5-8.

Tusinbaum H.A. and Ruvkun G.B. (1998). An insulin-like signalling pathway affects both longevity and reproduction in *C. elegans. Genetics* 148:703-717.

Velculescu V. E., Zhang L., Vogelstein B., and Kinzler K. W. (1995). Serial Analysis Of Gene Expression. *Science* 270:484-487.

Wadman M. (1999). Human Genome Project aims to finish 'working draft' next year. *Nature* 398(6724):177

Waterston R.H., Martin C., Craxton M., Huynh C.,Coulson A., Hillier L., Durbin R., Green P., Shownkeen R., Halloran N., Metzstein M., Hawkins T., Wilson R., Berks M., Du Z., Thomas K., Theirry-Mieg J. and Sulston J. (1992). A survey of expressed genes in *Caenorhabditis elegans. Nature Genet.* 1(2):114-123.

Waterston R. and Sulston J.E. (1998). The Human Genome Project: reaching the finish line. *Science* 282(5386):53-4.

Wilson R., Ainscough R., Anderson K., Baynes C., Berks M., Bonfield J., Burton J., Connell M., Copsey T., Cooper J. *et al.* (1994). 2.2Mb of contiguous nucleotide sequence from chromosome III of *C. elegans. Nature.* 368:32-38.

Wendl M., Dear S., Hodgson D. and Hillier L. (1998). Automated Sequence Pre-processing in a large scale sequencing environment. *Genome Res.* 8:975-984.

Wootton, J.C. and Federhen, S. (1996). Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.* 266:554-571.

Wolff A.P. and Brown D.D. (1988). Developmental regulation of two 5S ribosomal genes. *Science* 241:1626-1632.

Zhang H. and Blumenthal T. (1996). Functional analysis of an intron 3' splice site in *Caenorhabditis elegans. RNA* 2(4):380-388.

Zetka C-M. and Rose A.M. (1995). Mutant *rec-1* eliminates the meiotic pattern of crossing over in *Caenorhabditis elegans. Genetics* 141:1339-1349.